

The trials of gene therapy

The news that a child in a gene-therapy trial has developed cancer has cast a cloud over the technique. But this has more to do with the field's chequered history than the particular circumstances of this tragic case.

"Oh no, not again!" That was the reaction of many observers of gene therapy last month, after researchers in Paris announced that a child given a pioneering treatment for severe combined immunodeficiency disease (SCID) had succumbed to a leukaemia-like disease (see *Nature* **419**, 545; 2002).

Just three years ago, the death of 18-year-old Jesse Gelsinger, in a trial of a gene therapy for a liver defect that causes a dangerous build-up of ammonia, threw the field into crisis — with good reason. Gelsinger died of an inflammatory reaction to the viral vector used to deliver the corrective gene, and serious faults were soon found in the way that the trial, at the University of Pennsylvania in Philadelphia, had been conducted. Patients had been inadequately informed of the potential risks, despite evidence of problems at high doses of the vector in animal experiments. And when signs of liver stress emerged in some patients, regulators weren't informed.

As these facts sunk in, questions began to be asked about the reporting of adverse events in other trials, and gene therapists embarked on a bout of soul-searching. Perhaps, some acknowledged, the field had been too eager to rush into the clinic.

The present case, in which a retroviral vector's site of insertion into the genome seems to have activated a cancer-causing gene, has come as a further blow. And given recent history, it's understandable that authorities in France, directly responsible for the affected trial, and the United States, where the Gelsinger debacle has left painful memories, should move quickly to suspend such trials.

However, a closer examination of the French case reveals important differences to the Gelsinger case that should convince regulators to proceed once more with SCID trials, albeit more cautiously.

First, there has been no suggestion that the investigators are at fault. Cancer triggered by 'insertional mutagenesis' was always recognized as a risk of gene therapy using retroviral vectors, and the parents of patients enrolled in the SCID trials were informed of this possibility. The researchers involved are now pursuing studies to investigate the risks facing the other patients they have treated (see page 116). And although a definitive risk assessment isn't possible, the results should be promptly communicated to regulators and the patients' parents.

Perhaps the most important difference to the Gelsinger case is that, whereas SCID gene therapy is a potential life-saver for children who otherwise face an extremely bleak outlook, the Gelsinger trial was a safety study in adults of a treatment designed for young children. Gelsinger and the other volunteers did not stand to gain any therapeutic benefit. In trials of drugs for life-threatening conditions, on the other hand, severe adverse events do sometimes occur, and may be tolerated if the benefits outweigh the risks. It is in this light that the current SCID setback should be viewed.

This is not to say that changes are unnecessary. Procedures for obtaining informed consent from the parents of future SCID gene-therapy patients must be adjusted to stress that cancer caused by insertional mutagenesis is now a tangible, rather than a theoretical, risk. Regulators must also reconsider the wisdom of using retroviral vectors, particularly where genes are introduced to mark populations of cells for study, rather than for their intrinsic therapeutic effects.

The challenge for gene therapists and regulators is to show that the field can respond appropriately to a serious adverse event in an otherwise successful clinical trial. It is unlikely to be the last: such setbacks are inherent to the development of new medical treatments. ■

Frameworks can be too rigid

There are lessons to be learned from scientists' responses to the European Commission's latest research programme.

When asked about their experiences with the European Commission's Framework research programmes, scientists tend to roll their eyes. Baffling bureaucracy and the programmes' overt socioeconomic objectives have been deterring many of the continent's best researchers.

The Sixth Framework Programme, worth 17.5 billion euros (US\$17.7 billion) over the next five years, was highlighted this week with great fanfare at a major conference in Brussels. So is it likely to be any more popular than its predecessors?

The programme's underlying vision is an end to the fragmentation of European research. It aims to unite to an unprecedented degree the best groups in the fields that are most relevant to European citizens. Take breast-cancer research: the idea is that geneticists, molecular biologists, clinical researchers, biotech and pharmaceutical companies, and whoever else is relevant, will pool their expertise into an overarching research consortium with the goal of defeating the disease.

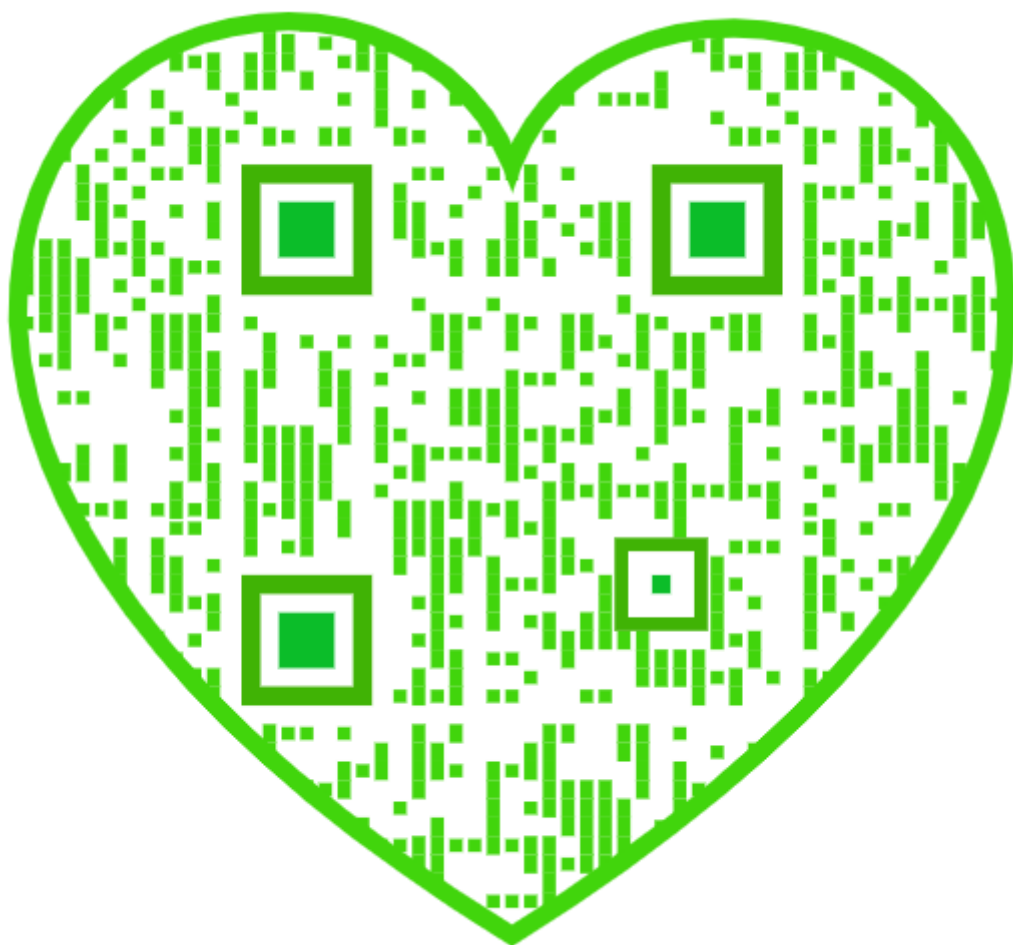
It sounds a laudable aim, but the concept fails to consider the role of competition as a driving force in science. Mega-collaboration is not always the best solution to a tough intellectual problem. And

the demands of managing huge research networks can detract from scientific creativity and hypothesis-driven enquiry.

Fortunately, however, the European Commission has received a great deal of indirect feedback about its plans. Rather than setting the programme in stone, and simply issuing a series of narrowly defined 'calls for proposals', the commission has this time invited scientists to submit 'expressions of interest' — condensed proposals outlining what researchers think are the most fruitful scientific opportunities (see http://eoi.cordis.lu/search_form.cfm).

The response was overwhelming, with some 15,000 proposals being submitted. These submissions do not just represent a goldmine of scientific ideas. Their sheer number, and the fact that many do not envisage huge collaborative efforts, also provides a clear indication that Europe's research community is not willing to accept the idea that a few large networks fit all.

How exactly this exercise in consultation translates into formal calls for proposals, scheduled for 17 December, remains to be seen. But it is not too late for the European Commission to adapt its vision, and accept that, in some cases, thinking small is the best approach. ■





On trial
Dismay greets plans
to test smallpox
vaccine on children
p110



Local uprising
Amerindians angry
over use of blood
samples in studies
p111



Weapon control
Royal Society backs
code of conduct for
biowarfare research
p112



Election blues
Scientists fret as
Republicans
celebrate
p113

Microbe hunt raises doubts over local benefits of bioprospecting

David Cyranoski, Tokyo

Controversy is likely to follow an international consortium of scientists — including researchers from the Chinese Academy of Sciences — as it embarks on a search for commercially useful microbes in the politically contested areas of Tibet and Inner Mongolia.

The scientists will be on the lookout for previously unknown extremophiles — microorganisms that live in inhospitable environments such as salt lakes and hot springs. But it is unclear if the two regions involved will benefit from any discoveries that are made.

The project is being led by microbiologist Bill Grant from the University of Leicester, UK, and includes scientists from the University of Seville in Spain, and the Netherlands branch of the US genomics company Genencor. Its main sponsor is the European Commission, which has donated 1 million euros (US\$1 million). The University of the Western Cape in Cape Town, South Africa, will bring its own funding to the consortium.

The researchers will isolate microbial DNA on site directly from the soil. They will sequence the DNA and search DNA databases for sequences encoding enzymes that give the microbes their special characteristics.

Grant and Genencor have already shown



An international consortium is seeking potentially valuable microbes in the harsh Tibetan landscape.

that the strategy can be commercially successful. A collaborative expedition to Kenya in 1992 found an extremophile in a soda lake that has an enzyme that breaks down cellulose over an unusually wide range of temperatures. In 1998, Genencor exploited this property in a process for stonewashing jeans.

China hopes to gain commercially from the project. Its science ministry has negotiated with the consortium to retain “sovereign rights” to biological resources found and a share in any commercialization, says Yanhe Ma of the Chinese Academy of Sciences Institute of Microbiology in Beijing, a member of the consortium. But there is no indication

so far that a fair share of the benefits will filter back to Tibet and Inner Mongolia — contentious “autonomous regions” of China — or even that the issue has been discussed. China’s autonomous regions are supposed to retain a degree of control over local economic interests, but many commentators argue that Beijing has neglected these rights.

The United Nations’ Convention on Biological Diversity, which the European Commission and China signed in 1992, aims to ensure that countries, particularly indigenous and local communities, benefit fairly if their biodiversity proves to be commercially valuable. Ma says that local governments will be helping them with the project, but he gives no details of any plans to bring work or other benefits to the local people.

Huanming Yang, director of the Beijing Genomics Institute and an advocate of the protection of China’s genetic resources, says that scientists from local communities should be involved. Chinese researchers should be sensitive to such issues, he says, but foreign researchers should also bear responsibility. John Ackerly, president of the Washington-based human-rights group International Campaign for Tibet, agrees. “It’s the responsibility of the European researchers to make sure that Tibetans are engaged and given benefits, because Beijing won’t do it,” he says.

But this is not an easy task. “How could an outsider possibly go about pushing the issue of regional recognition?” asks consortium member Don Cowan of the University of the Western Cape.

Ground-breaking web test gears up

Geoff Brumfiel, Washington

Earthquake engineers are eagerly anticipating the first test of a national network of seismic experiments. The George E. Brown Jr Network for Earthquake Engineering Simulation (NEES) is an \$82-million project, sponsored by the National Science Foundation, to build a series of 16 centrifuges, shake tables and other large-scale experimental equipment at universities across the United States.

When they are completely integrated in 2004, these earthquake simulators will be part of a network of experiments and software tools that will enable engineers to

conduct their research through the web.

The first test of the web-interface technology takes place this week at the University of Nevada in Reno. A shake table will vibrate a model bridge fitted with about 100 sensors. These sensors will stream video and data straight into the computers of watching engineers, who will then attempt to analyse the bridge’s performance.

“This is the first public peek at this infrastructure,” says Dan Reed, director of the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, who oversees the computing aspects of NEES.

Call for more data forms basis of Bush climate strategy

Tony Reichhardt, Washington

Improved computer models, accelerated studies of the global carbon cycle, and standardization of climate data are among the priorities listed in the Bush administration's draft strategy for climate-change research, released on 11 November.

The plan sets an overall direction for the Climate Change Science Program, which was created last year to address near-term uncertainties in climate science. The new programme merged the \$1.7-billion, multi-agency US Global Change Research Program (USGCRP) with the narrower \$40-million Climate Change Research Initiative.

The new strategy acknowledges that global warming is under way, but reflects the Bush administration's wariness on the subject by emphasizing gaps in knowledge that need to be filled before policy-makers can make confident decisions.

As well as listing high-priority research needs, such as characterizing atmospheric aerosols, the plan calls for investment in observational networks, including an Integrated Ocean Observing System.

The strategy will now undergo extensive review, starting with a workshop in Washington next month for scientists and policy-makers. It also will be vetted by a fast-track National Academy of Sciences panel chaired by Thomas Graedel, director of Yale University's Center for Industrial Ecology. A final version of the plan is expected to be released next April.

The plan's authors say that they relied heavily on previous academy reports that outlined priorities for climate-change research. But conspicuously absent from the document is any mention of the controversial 'National Assessment' published by the USGCRP in 2000, which described possible regional effects of climate change in the United States. The new strategy states flatly that modelled projections "are often contradictory and are not sufficiently reliable tools for planning".

The National Assessment infuriated global-warming sceptics, who sued to prevent its release, charging that it was politically motivated. The anger has not subsided. Last month Christopher Horner, a senior fellow at the Competitive Enterprise Institute (CEI), a public-policy organization based in Washington, called for members of the team that produced the National Assessment to be barred from next month's workshop. ■

► www.climate-science.gov



A. REEVE/SPL

Many critics argue that using small children in a trial for a smallpox vaccine poses too greater a risk.

Outcry greets US plan to test smallpox vaccine on children

Erika Check, Washington

Plans to test a smallpox vaccine on small children in the United States have been slammed by critics who claim that the trial is unnecessary and potentially harmful.

The government wants to give 40 children, aged 2–5, doses of the Dryvax smallpox vaccine. Half would get a diluted form of the vaccine, and the rest would receive the full-strength version. The study, which would be funded by the National Institute of Allergy and Infectious Diseases in Bethesda, Maryland, would aim to assess the safety and efficacy of the diluted vaccine stockpile.

But the planned trial was greeted with a flood of criticism when details were posted on the website of the Food and Drug Administration (FDA) on 31 October as part of the public-review process.

The Department of Health and Human Services (DHHS) convened a panel of 10 experts — clinicians, lawyers and pharmacists — to examine the ethical questions raised by the proposed study. In their comments, which are also posted on the FDA's website, most of them agreed that there is a case for going ahead with the trials.

The study raises important questions, says Mary Faith Marshall of the University of Kansas Medical Center, who is former chair of the now-disbanded National Human Research Protections Advisory Committee and a member of the expert panel. For instance, she says, the risk of a smallpox attack is low, and an attack would probably affect just a few geographical regions. So most of the children who participate in the study will probably never be involved in an attack — and so would be unlikely to benefit from the study. On the other hand, between two and ten children out of every million could develop life-threatening side-effects from the vaccination.

Marshall points out, however, that the panel of experts recognized that the risk of an attack is not zero. And a smallpox attack would be deadly, killing about 30% of infected individuals. "That argues more strongly than if the mortality was not as high," says Marshall.

But others disagree with the panel members' general approval of the proposal. They point out that despite the Bush administration's assertions that Iraq has stockpiles of smallpox virus, there is still no proof that it or other nations are planning to attack the United States with germs. And, they say, scientists are already developing a safer version of the vaccine, which will probably need to be tested in children anyway.

"The risks of the Dryvax vaccine are clear — and the fact is that death can occur even in normal, healthy children," says Paul Offit, chief of infectious diseases at the Children's Hospital of Philadelphia in Pennsylvania and a member of the Advisory Committee on Immunization Practices, which advises the DHHS on vaccine policy. "There could be children who suffer from the vaccine who never had any chance of suffering the disease."

And Al Sommer, dean of the Johns Hopkins Bloomberg School of Public Health in Baltimore, Maryland, points out that young children are not at high risk of contracting smallpox in the event of a bioterrorist attack. Also, he notes, earlier experience with the vaccine indicates that young children are unlikely to respond to the vaccine differently from adults. "None of this makes any sense from a public-health perspective," Sommer says.

The FDA is accepting public comments on the smallpox trials until 2 December. The federal health secretary and the head of the FDA will then issue a final decision on the trials. ■

Tribe blasts 'exploitation' of blood samples

Rex Dalton, Vancouver

Nearly 20 years ago, Amerindians on Vancouver Island in Canada donated blood for research into the genetic causes of rheumatoid arthritis, a disease that is rampant in their tribe. Now they are incensed to discover that the specimens have been used for other research — including a project on the sensitive issue of the spread of lymphotropic viruses by intravenous drug abuse.

Leaders of the Nuu-chah-nulth (Nootka) tribe describe the research as another example of exploitation of indigenous peoples. Some have demanded the return of the samples. The case has opened a debate on what should happen to stored biological samples from completed projects, and whether they should be allowed to be moved between different research institutions.

The arthritis research was conducted at the University of British Columbia (UBC), and led by geneticist Ryk Ward, who collected the blood samples and took genealogical histories from the tribe. Ward left UBC in 1986 and took the Nuu-chah-nulth specimens with him, first to the University of Utah and then, in 1996, to the University of Oxford, UK, where he continues to head the Institute of Biological Anthropology. Unable to show a genetic basis for the tribe's arthritis with the tools available in the 1980s, Ward used the samples for other research projects, sharing data with many different collaborators. These projects yielded half a dozen published articles.

UBC officials confirm that the tribal participants did not give consent for research on subjects other than arthritis. Ward also acknowledges that he never obtained renewed consent from the tribe to take the blood samples with him for further studies. "The way people operated at the time," he told *Nature*, "it didn't cross anyone's mind — we didn't mean to be evil, and we are more careful now."

The genetics community is becoming more sensitive to the issue, and some institutions in Canada and the United States have recently tightened rules on consent and the use of stored samples. "But it remains a major concern," says Michael McDonald, a UBC ethicist who advises Canadian research-funding agencies. "Researchers ship biological specimens all over, but we don't have good procedures for dealing with this."

In the wake of the Nuu-chah-nulth case, UBC and Utah have introduced policies that require researchers to obtain consent each time they wish to conduct new research on stored samples. And last January, Utah implemented another policy whereby study samples were declared to be university property, and cannot be removed by departing researchers without specific permission. "It is inappropriate for folks to drive away with biological samples," says geneticist Jeffrey Botkin,



Members of Canada's Nuu-chah-nulth tribe are angry at the unsanctioned use of blood samples.

Utah's associate vice-president for research.

Laura Arbour, a UBC geneticist, is writing a case study on the Nuu-chah-nulth experience. The problem with genetic samples, she says, is that the research community has traditionally considered the DNA as a gift, when

she says it really is only "loaned" under a very specific contract. And genetic research on Amerindians — who are much sought after for their distinctive genetic lineages — is even more sensitive, Arbour notes.

The Nuu-chah-nulth problem came to the attention of UBC authorities two years ago, after an indigenous peoples' newspaper reported tribal unrest over the use of their samples for non-arthritis research. Canadian geneticists and government officials have since held five workshops to review the case, and to develop better policies.

Among the proposals under consideration is a national tissue bank for genetic samples from First Nation tribes, who would govern the repository themselves. Some Canadian leaders question the practicality of such an enterprise, as it would be difficult to get the diverse tribes to agree a strategy.

International rules are needed, says McDonald, to ensure that human biological samples are respected — particularly now that the human genome has been sequenced, tempting more scientists to reach into freezers for stored samples for comparative analyses. ■

Space station set for sky-high profile

David Adam, London

Already faced with cost overruns and technical delays, officials overseeing the building of the International Space Station (ISS) are scratching their heads over a new problem — determining its 'values'. Desperate to boost funds for the project, the European and Canadian space agencies have enlisted the help of a public-relations company to develop a "branding and communications strategy" for the cash-strapped station.

"In the past, human spaceflight has wrongly retained an exotic and élitist image. With the ISS this is changing," explains Jörg Feustel-Büechl, director of manned spaceflight and microgravity at the European Space Agency (ESA). "The station offers companies and organizations the opportunity, for the first time, to exploit manned spaceflight for commercial purposes."

As well as focusing on the research and development community — many of whom, ESA says, are unaware of the station's potential — the agency is also targeting what it calls "newer kinds of space entrepreneurs" involved in marketing, entertainment and tourism.

Still, those looking at the station are unlikely to see a cola advert painted onto its side, says Nicholas Lunt, business development director with Ogilvy Public

Relations, the Brussels-based firm hired by ESA. "If you view this solely as painting logos on pieces of hardware then that misses the point of what sponsors are looking for," Lunt says. "We cannot pretend in any way that this is a Formula One car."

The firm — which Lunt says is a "360-degree communications agency" — has signed a one-year deal to determine the brand position of the ISS, and to develop a communications strategy to put it into the public eye and raise its profile among the business community. "And we can't do that until we've identified the values of the station," he says. ■



Conduct code mooted for bioweapons treaty

Natasha McDowell, London

A code of conduct for scientists could help to strengthen the international Biological and Toxin Weapons Convention (BTWC), according to Britain's Royal Society.

Negotiations aimed at reviewing the 1972 convention broke down last December, when the United States blocked plans to allow laboratory inspections as a means for checking compliance with the BTWC (see *Nature* **414**, 675; 2001).

The British government floated the idea of a code of conduct earlier this year as a way to strengthen the BTWC even if the United States maintains its position. The Royal Society's endorsement of the proposal, contained in a report released on 6 November, comes as negotiations on the convention resume this week in Geneva.

The report argues that there is considerable ignorance of the BTWC among British researchers, and that a code of conduct would encourage all scientists to consider the biowarfare applications of planned research.

The society did not give details of a code, but noted that some institutions have already developed their own rules. Some, for example, require researchers to assess biowarfare applications of their work formally before beginning a project.

The need for such formal assessments

was highlighted last January, when Australian researchers working on a contraceptive vaccine for rodents inadvertently created a highly virulent strain of mousepox (see *Nature* **411**, 232–235; 2001).

Brian Spratt, a molecular biologist at Imperial College in London and a member of the committee that compiled the report, believes that the code of conduct could be incorporated into existing mechanisms. Many universities have safety committees that assess the risk of experiments such as creating strains of genetically modified crops, for example. Spratt says that the same committees could also consider biowarfare implications of proposed research. A professional body within each country with the ability to dole out sanctions could be another enforcement mechanism, the report's authors say.

In addition to the code of conduct, the report recommends that biowarfare issues should be introduced into academic courses. And it calls for an international scientific advisory panel to monitor new developments regularly, in an effort to keep up with the rapid progress in biological technologies.

"It would mean that advances could be kept under international review in a more systematic way than at present," says Brian Eyre, a materials scientist at the University of Oxford, who chaired the committee. ■



Pillars of society: the Royal Society wants to see education about potentially dangerous research.

ROYAL SOCIETY

Academy slams Internet arts and sciences lookalike

David Adam, London

Researchers may do well to heed Groucho Marx's comment: "I don't want to belong to any club that will accept me as a member."

Two weeks ago *Nature* revealed that the self-styled European Academy of Sciences appears to have no high-brow status (see *Nature* **419**, 865; 2002), despite grandiose claims made to scientists invited to become members. Now, it seems, a similar academy has emerged across the Atlantic.

Over the past few months, the North American Academy of Arts and Sciences — not to be confused with the American

Academy of Arts and Sciences, which is based in Cambridge, Massachusetts — has bulk-distributed e-mails to researchers, inviting them to become members.

Membership and its 'benefits' — such as the opportunity to publish in the academy's journals — are free, although members are encouraged to send off \$20 for a certificate. On its website the academy claims to offer scholarships and awards, and to organize conferences and contests. It laments that "all too often, those who make the greatest contribution to our society go unrecognized".

But *Nature* has been unable to find any record of the academy's publications, awards or conferences, or to contact its organizers. No names appear on any of its e-mails or on its website, which is registered to a postal address in Budapest, Hungary. Enquiries sent to several of the academy's listed e-mail addresses went unanswered.

The American Academy of Arts and Sciences says it has had the website closed down several times, only to see it spring up again. It is now considering legal action.

Many scientists have become members of the self-styled "learned society", according

to its website, joining a feng shui expert, a hypnotherapist and a town crier. Computer scientist Kia Ng, of the University of Leeds, UK, says he knows little about the academy, but joined "because it only involved clicking a link". He has not bought a certificate.

"It is nice to say you are a member, especially if you don't pay anything," says Dragan Ciscic, a researcher in maritime studies at the University of Rijeka, Croatia.

Others say that their names are being used without their knowledge. Hossein Arsham, a statistician at the University of Baltimore, Maryland, is listed as a member, but says that he has never heard of the academy and that its information about him is years out of date.

The academy offers its fellows the opportunity to make money by writing exam questions for an online certification programme, to be offered through the "New York College of Advanced Studies". The college's website, which gives no details of its accreditation or organizers, says that for "three easy payments of US\$21", students will be able to take tests and receive certificates from the college, described as an affiliate of the academy. ■

North American Academy of Arts and Sciences

WELCOME

Congratulations on being recommended to be a fellow of the North American Academy of Arts and Sciences. This is a great honor and a testament to your achievements in your field. We are pleased to have you join our ranks of distinguished members.

WELCOME OUR NEW FELLOWS

We are pleased to have you join our ranks of distinguished members. We are committed to providing you with the resources and support you need to succeed in your field. We are also committed to promoting the highest standards of research and scholarship.

Primary Policy

The primary policy of the North American Academy of Arts and Sciences is to promote the highest standards of research and scholarship. We are committed to providing our members with the resources and support they need to succeed in their field. We are also committed to promoting the highest standards of research and scholarship.

Academy awards: "If you choose to accept this invitation" — you will be in colourful company.

Republican win sparks fears for science funds

Geoff Brumfiel, Washington

With the Republicans wresting control of the US Senate from the Democrats in last week's mid-term elections, science leaders are starting to worry about the outlook for science funding in the United States.

Fears are growing that the Republican's dominance of both houses in the Congress will spell trouble for the future budgets of key science agencies. Prominent scientists are also worried that there will now be delays in the payment of grants and funds for building new facilities.

The priority for the Republican leadership over the next two months is the creation of the Department of Homeland Security. This means that long-standing disagreements over the 2003 budget are likely to wait until the new year.

"This will create some real problems for science," says Mike Lubell, director of public affairs at the American Physical Society. This December, the National Institutes of Health (NIH) was set to begin handing out grants from a massive \$3.7-billion boost it had expected to receive as the final instalment of a five-year budget-doubling process. And the National Science Foundation (NSF) was hoping to start building several projects and facilities that may now be delayed.

A bill that would have recommended doubling the NSF's budget over the next five years may also be in peril because the White House is reportedly not in favour of it (see *Nature* 419, 657; 2002).

The Republican takeover of the Senate will also stoke the debate over therapeutic cloning, says Kevin Wilson, director of public policy at the American Society for Cell Biology. In July of 2001, the House of Representatives passed a bill banning reproductive and research cloning. Since then the Senate has been debating whether cloning for research purposes should be allowed. That debate may shift under the leadership of Trent Lott (Republican, Mississippi), who is expected to take over as Senate Majority Leader, and who has opposed cloning in the past.

But a total ban is thought to be unlikely. "There are still some conservative Republicans who support research cloning," says Joanne Padrón-Carney, director of the Center for Science, Technology and Congress at the American Association for the Advancement of Science. Two such advocates, Arlen Specter (Republican, Pennsylvania) and Orrin Hatch (Republican, Utah), will chair committees that could prove influential in the debate.

Wilson still believes that the research community must redouble its efforts to convince Congress and the Bush administration of the virtues of research cloning. "Last year there was a threat of a comprehensive cloning ban," he says. "That threat



Victory: Republicans celebrate winning control of Congress in the mid-term elections.

has become a little more real to us."

The shift of power may signal difficult years ahead for research funding, scientists fear, as the recession and a possible Iraq con-

flict push research further down the priority list. "In the coming year," says Lubell, "I think the discretionary budget for science is going to be exceedingly tight."

Venus Express cleared for take-off

Alison Abbott and Quirin Schiermeier

Europe's mission to Venus is no longer in any doubt. On 5 November, the European Space Agency (ESA) confirmed that Venus Express, scheduled for launch in 2005, will go ahead.

The mission to Earth's hotter neighbour was cancelled in May because of funding delays (see *Nature* 417, 474; 2002), but the decision caused such alarm in the space-science community that the agency rethought its plans. In July, ESA began work on the design phase of the mission, in anticipation of this month's final approval.

The fate of Venus Express had been in doubt because Italy could not finance one of the mission's most important instruments, the Visible and Infrared Thermal Imaging Spectrometer (VIRTIS), being developed at the Institute for Space Astrophysics in Rome. This instrument will detect the low levels of near-infrared radiation that escape Venus's thick atmosphere, and so allow scientists to study the composition of the planet's lower atmosphere.

ESA has now agreed to cover the costs of integrating the instrument into the spacecraft, even though it normally expects national space agencies to pay for scientific instruments. In return, ESA will publish a call for researchers from other member

states to join the VIRTIS team.

"VIRTIS is a crucial instrument for the mission, and that's why we fought so hard for it," says a relieved Dmitri Titov, the mission's principal investigator, who is based at the Max Planck Institute for Aeronomy in Lindau, Germany.

But other projects were not so fortunate. DIVA, a planned German astrometry satellite to measure the position and movements of 35 million stars in the Milky Way, and planned for a 2004 launch, now looks unlikely to take off.

Germany had asked ESA to help close a 15-million-euro (US\$15.2-million) funding gap. But the agency's scientific advisers said that participation in DIVA could distract European astrometrists from ESA's cornerstone astrometry mission, GAIA, scheduled for launch in 2012.

Siegfried Röser, DIVA's principal investigator and a senior scientist at the Astronomisches Rechen Institute in Heidelberg, argues that DIVA would have been "an ideal precursor for GAIA in testing technology". But unless NASA accepts a proposal currently being considered by the US Naval Observatory in Washington DC for a US-German partnership on the mission, the ESA decision is the death knell for DIVA.

B. ALKOFER/CETTY IMAGES

Developing-world centres plan for farm funds to flood in

Tokyo Schemes aimed at wooing new financial donors for agricultural projects have been launched by the Consultative Group on International Agricultural Research (CGIAR), a global network of 16 research centres that address the farming needs of poor countries.

CGIAR hopes that new water-management and nutrition projects will bring money into the group, and reverse a trend that has seen some cash-strapped centres forced to lay off staff (see *Nature* **416**, 777; 2002). The water-resources project will involve using biotechnology and traditional breeding techniques to boost agricultural yields without increasing water usage. New breeds of rice, which can grow in non-flooded conditions, are one possibility. Methods for increasing the amount of iron, zinc and vitamin A in key crops such as rice, wheat and maize will be studied in the nutrition programme.

CGIAR wants to raise a total of US\$130 million for the first five years of what it hopes will develop into programmes that last at least ten years. The World Bank has already announced that it will support both

programmes, and the Dutch government has agreed to donate 25 million euros (US\$25 million) to the water project.

Darwin's finches threatened by bloodthirsty invaders

London The Galapagos Islands' unique finches, which helped Darwin to formulate his theory of evolution, are under threat from an invasive species of parasitic flies.

The flies probably arrived in a shipment of fruit or vegetables. Their larvae live in the finches' nests, burrowing into nestlings and drinking their blood. Ornithologists Birgit Fessl and Sabine Tebbich of the Konrad Lorenz Institute in Vienna monitored the flies between 1997 and 2000. They report that around a quarter of nestlings in infected



Under attack: Darwin's finches, a model of evolution, must adapt to life with parasitic flies.

nests die (B. Fessl and S. Tebbich, *Ibis* **144**, 445–451; 2002).

Fessl and Tebbich say that insecticide can be used to kill the larvae. But Nigel Collar, a specialist in American birds at the conservation group BirdLife International in Cambridge, UK, says that the best hope is for the finches to learn to live with the flies. "You can't go around fumigating every Darwin finch nest for the rest of time," he points out.

Indian to boost physics in poorest countries

Rome Indian-born experimental physicist Katepalli R. Sreenivasan has been named director of the Abdus Salam International Centre for Theoretical Physics in Trieste, Italy. The United Nations-backed centre, which was founded in 1964, trains physicists and fosters research in the developing world.

Sreenivasan received a doctorate in aerospace engineering from the Indian Institute of Science in Bangalore in 1975. He subsequently worked in Australia and the United States, conducting research in fields such as complex fluids, combustion, cryogenic helium and nonlinear dynamics.

Sreenivasan is scheduled to take over from current director Miguel Angel Virasoro

next March. Virasoro's appointment in 1995 was marred by complaints that a researcher from a developing country should have been given the post. Virasoro is Argentinean and has worked in Italy for most of his life.

♦ www.ictp.trieste.it

Snappy theft from van robs science of crocodile fossil

Johannesburg A 250-million-year-old amphibian fossil skull is missing, presumed lost, after thieves stole the parked van it was left in overnight. The fossil was the only one of its type, and the specimen on which the name and first description of the crocodile-like *Jammerbergia mucidiops* was based. It belonged to the National Museum in Bloemfontein, South Africa, and was being transported back there when the thieves struck late last month.

Ross Damiani, a researcher at the Bernard Price Institute for Palaeontological Research at the University of the Witwatersrand in Johannesburg, describes the theft as a "huge loss". He had borrowed the fossil to write a research paper on it, which is due to be published in the *Journal of Vertebrate Paleontology* next year. Police say it is very unlikely that either the vehicle or the fossil will be recovered.

Wellcome approval for Hinxton expansion

London After five years of protests and planning wrangles, the Wellcome Trust last week got the green light to expand its Genome Campus in Hinxton, Cambridgeshire. The local council gave its approval on 6 November for the scaled-down extension, which will carry out postgenomic research and development of new medicines and therapies.

The trust, which is Britain's leading biomedical charity, had its previous application for a 40,000-square-metre extension turned down in 1999 (see *Nature* **400**, 803; 1999). It resubmitted plans for the 27,000-square-metre development — which no longer included space for established biomedical companies — earlier this year.

"We can now look forward to playing an instrumental role alongside industry to develop new medical advances," says Allan Bradley, director of the Wellcome Trust Sanger Institute, which is based at the campus.

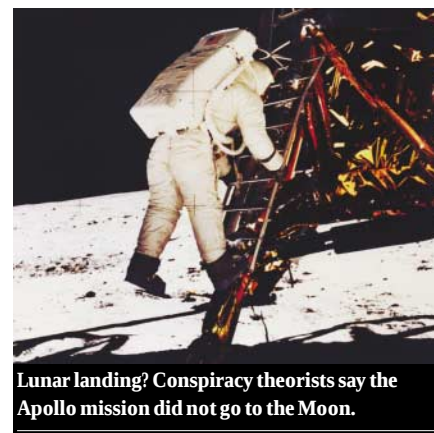
One small step backwards as NASA scraps hoax denial

Washington Could it be a giant leap for conspiracy theorists? Embarrassed by what it saw as negative publicity, NASA last week scrapped a planned monograph that would

have presented arguments against the claim that the Apollo moon landings were a hoax (see *Nature* **420**, 9; 2002).

NASA spokesman Bob Jacobs says that the original intent of the project was lost in news reports that portrayed the plan as an attempt to convince the sceptics, rather than inform teachers and the public through a "serious academic research project".

An ABC News television broadcast on 4 November rankled more than other coverage. In it, anchorman Peter Jennings said: "A professor of astronomy in California said he thought it was beneath NASA's dignity to give these people the time of day. We simply wonder about NASA."



Lunar landing? Conspiracy theorists say the Apollo mission did not go to the Moon.

NASA



Better days: Alain Fischer and Marina Cavazzana-Calvo announce the successful results of their gene-therapy trial on immune-deficient children in April 2000.

A tragic setback

With one French gene-therapy patient having developed a form of cancer, a frantic detective effort is under way to determine what went wrong — and to assess the risks faced by others. Erika Check reports.

Until a few weeks ago, a three-year-old boy whose identity remains confidential was a beacon of hope for gene therapists. He is one of 11 children with severe combined immune deficiency (SCID) to receive a pioneering treatment from a team led by Alain Fischer and Marina Cavazzana-Calvo of the Necker Hospital for Sick Children in Paris. It worked, providing the first proof that gene therapy can cure a life-threatening disease¹.

Now, however, the boy has developed a leukaemia-like condition that most experts believe was caused by the very treatment that cured his SCID. The field of gene therapy is in turmoil, as scientists and medical authorities try to figure out how to proceed with trials in SCID and other diseases². For Fischer, the past few weeks have been a blur of meetings, e-mails and telephone calls, as he has kept colleagues informed of developments. But already his work has begun to reveal what

went wrong in this tragic case. And future findings may allow an assessment of the risks faced by other patients.

The 11 children in Fischer and Cavazzana-Calvo's trial, 9 of whom have been cured, suffered from a form of SCID caused by a defect in a gene on the X chromosome. The gene encodes a subunit of a cell-surface receptor that allows developing immune-system cells to respond to growth signals called cytokines. Without this subunit, known as the γ -c chain, children fail to develop the mature T cells, B cells and natural killer cells needed to fight off infections. The condition can be cured by a perfectly matched bone-marrow transplant, but few children have perfect donors. Of those who receive imperfectly matched transplants, up to 30% die.

The French trial offered new hope to children with the disease by replacing the defective gene. Doctors at Great Ormond Street

Children's Hospital in London have since treated four patients. Gene therapists in the United States and Italy³ have treated another form of SCID, caused by a defect in the gene for the enzyme adenosine deaminase (ADA), which is needed for immune-cell development. And another US group has treated one patient whose SCID results from a deficiency in a gene called *JAK3*, which produces a protein that is involved in transmitting the signals received by cytokine receptors.

In each of these trials, the corrective genes have been packaged into modified retroviruses, which can incorporate themselves into a host cell's DNA. These retroviruses have been stripped of most of their viral genes to prevent them from causing dangerous infections, but gene therapists cannot control where retroviral vectors insert themselves. So one long-standing worry is that a vector could disrupt important genes through 'insertional mutagenesis' — and if

the gene involved is one that normally regulates cell growth and division, cancer could result.

This seems to be what happened in the case of the French boy, who has developed a leukaemia-like condition in which one particular type of T cell began to proliferate uncontrollably. As a result, SCID gene-therapy trials in France and the United States have been suspended, Italian authorities have suspended the enrolment of new patients, and planned trials in Japan have been put on hold. Germany, meanwhile, had already suspended a range of gene-therapy trials involving retroviral vectors in the wake of a report⁴ that a such a vector has caused leukaemia in mice.

But given that many patients have been treated in gene-therapy trials involving retroviral vectors, and there is no other known instance of vector-triggered cancer, most experts believe that the benefits still outweigh the dangers. Britain's authorities, for instance, did not suspend trials in light of the French setback. And an advisory committee to the US Food and Drug Administration (FDA) has now advised that trials under the agency's jurisdiction should resume. "None of these types of adverse events has ever been seen before," says Savio Woo, a gene therapist at Mount Sinai School of Medicine in New York. "The frequency is really very low."

But other experts argue that it's too early to know how great the risk is. The patients exposed to retroviral gene-therapy vectors were all treated in the past few years. So it is possible that some will develop cancer at some time in the future. This is why Fischer is now working to determine what, exactly, went wrong in his unfortunate patient — and whether any other children face a similar fate.

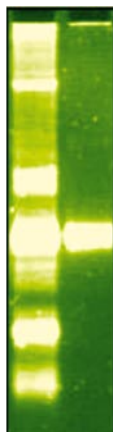
To do so, Fischer has enlisted the help of molecular biologist Christof von Kalle of the Cincinnati Children's Hospital. While at the University of Freiburg in Germany, von Kalle and his colleague Manfred Schmidt developed a technique called linear amplification-mediated PCR (LAM-PCR) to detect where a retroviral vector inserts into human DNA⁵.

LAM-PCR is a variant of the standard polymerase chain reaction, used to pull out a specific DNA sequence from a sample of DNA. PCR requires a short 'primer' sequence that binds tightly to part of the sequence being sought. Using a primer for the retroviral vector, therefore, it is possible to use PCR to selectively amplify the vector sequence, if it is present, and so to detect whether the vector has incorporated into a particular cell or tissue. But that still leaves the question of where in the genome it is integrated.

Von Kalle and Schmidt's technique uses a



Christof von Kalle (above, left) has devised a way to track the insertion sites of gene vectors. Successive analyses can highlight the potentially cancerous transition from a healthy T-cell population with a range of insertion sites (left, left column) to one dominated by a single site.



retroviral vector primer tagged with a molecule called biotin, which binds tightly to another molecule called avidin. Performing PCR and washing the DNA sample over a column of avidin beads pulls out amplified DNA that contains the retroviral vector. The amplified sequences are then cut into pieces using a restriction enzyme and small 'linker' pieces of DNA with a known sequence are attached to the ends of the resulting fragments. Using these linker sequences as primers for further PCR, it is then possible to amplify the DNA that flanks the inserted vectors, so that it can be analysed to determine where the vectors have landed.

When Fischer and von Kalle used LAM-PCR to analyse the French patient's cancerous T cells, they found that the vector had inserted itself into a gene called *LMO-2*, mutations in which are known to be involved in childhood cancers. What's more, von Kalle found that the child's T cells were expressing *LMO-2*, which is abnormal. This has convinced most experts that the powerful promoter sequence in the vector, included to boost expression of the corrective gene, acti-

vated *LMO-2*, triggering the child's cancer. "Altogether, I think insertional mutagenesis is the most likely explanation," Fischer says.

Now, Fischer wants to use LAM-PCR to determine whether such problems are likely to occur in future, by making a catalogue of vector-insertion sites in SCID gene-therapy patients. So far, he and von Kalle have examined two other patients from the French trial. In total, they have now identified more than 100 insertion sites. None of those examined so far are in genes that scientists think are involved in cancer.

As von Kalle and Fischer fill in their insertion map, they may also find more information about where retroviruses prefer to insert themselves. In August, Frederic Bushman, a molecular virologist at the Salk Institute for Biological Studies in La Jolla, California, reported that HIV-1 inserts more often into genes than into non-coding regions of DNA⁶. The SCID trials involve a different virus, called the Moloney retrovirus. But if it, too, shows a penchant for inserting into genes, it will mean that the risks associated with SCID gene therapy are greater than was thought.

Identifying every single insertion site in a child's cells is an arduous process. Even this may not yield a complete picture of the risks that the children face, because promoter sequences can influence the expression of genes at remote sites in the genome. And in any case, we don't yet know the identity of all of the genes that can contribute to cancer.

Carrying out successive LAM-PCR analyses over time on individual patients could, however, tip gene therapists off to a developing leukaemia-like condition. Leukaemia occurs when an individual immune cell starts proliferating wildly so that, over time, it dominates the population

Given that there is no other known instance of vector-triggered cancer, most experts believe that the benefits still outweigh the dangers.

of cells circulating through the body. In LAM-PCR, this would show up by a patient initially carrying T cells with a range of different insertion sites. But over time, T cells with one particular insertion type would become more and more common until they were completely dominant.

Looking back at blood samples taken from the French boy, von Kalle has found that this is exactly what happened. Early on, the patient's T cells carried about 50 different insertion sites. But eventually, all of his T cells were clones of one cell with the LMO-2 insertion.

Gene therapists would like to include LAM-PCR analysis as a monitoring tool in SCID trials. But they note that not all children with a single insertion site have leukaemia. Paediatrician Donald Kohn of the Children's Hospital in Los Angeles, who is running a trial on ADA SCID, treated one patient nine years ago whose T cells all carry the same insertion site. Yet this patient appears completely healthy. Kohn believes that, in this case, the vector integrated into a bone-marrow stem cell that has divided normally to derive the child's entire T-cell population, rather than there being a cancerous proliferation of a particular T cell. For this reason, Kohn says that analyses of insertion sites must be combined with detailed clinical information.

Such in-depth monitoring will help scientists to understand more about the risks of insertional mutagenesis in SCID gene therapy. But there is still a host of unanswered questions about other factors that might have predisposed the French child to cancer. A phenomenon called 'selective advantage' could be at work. As SCID patients are born with defective immune systems, only cells that integrate the corrective gene will survive. And these cells must divide millions of times to create a functional immune system. This could boost the chance that a cell will develop new harmful mutations, and could also increase the risk that mutated cells will thrive.

Then there are factors that are particular to the French patient's case. Two of his relatives had childhood cancers, which may indicate an inherited susceptibility. And he suffered a bout of chicken-pox before his immune system had fully developed — stimulating his immune system in a way that may have triggered T-cell proliferation. Matters are further complicated by the fact that the boy has a form of cancer that has never been seen before. His T cells, for instance, do not produce typical leukaemia proteins. "It's not at all clear what this disease is," says Ted Friedmann of the University of California, San Diego, who chairs the US National Institutes of Health's Recombinant Advisory Committee (RAC), which reviews federally funded gene-therapy trials.

All of this makes it difficult to draw general conclusions from the case. "There are a lot



Donald Kohn says that analyses of insertion sites must be paired with detailed clinical information.

of questions about the degree to which we can extrapolate to other trials," says Amy Patterson, head of the office that supports the RAC. But at the RAC's next meeting, on 4 December, it will have to begin wrestling with the issues.

One question that is likely to come up concerns the safety of 'gene-marking trials' in which doctors use retroviral vectors not to correct a genetic defect, but to introduce a marker gene to track a particular group of cells as they migrate through the body. For example, in patients given bone-marrow transplants, the introduced cells can be tagged with a genetic marker so that doctors can follow them and determine whether they develop normally. Patterson says that 41 marking trials are currently under way in the United States. But in the light of the French patient's cancer, some experts are wondering whether such trials should be discontinued.

Vector calculations

In the long term, gene therapists must consider whether the risks posed by insertional mutagenesis should cause them to move away from current retroviral vectors. They might, for instance, abandon retroviruses in favour of vectors that do not integrate into the genome. These aren't devoid of problems — as demonstrated by the 1999 death of teenager Jesse Gelsinger, killed by an inflammatory reaction to an adenovirus vector in a gene-therapy trial at the University of Pennsylvania in Philadelphia. But adenoviral vectors have since been refined to make them less likely to provoke dangerous immune reactions. Other non-

integrating vectors, such as modified Epstein-Barr and herpes viruses, are also being considered.

But it should be possible to engineer retroviruses to reduce the risks posed by insertional mutagenesis. One obvious possibility is to equip the vectors with 'suicide' genes. For instance, a vector could include a gene for a protein that renders any cell that produces it susceptible to the antiviral drug gancyclovir. If doctors noticed that cells that have integrated the vector were growing out of control, they could then administer the drug and nip the cancer in the bud.

Such approaches are already being used experimentally to attack difficult-to-treat cancers. Dan Salomon of the Scripps Research Institute in La Jolla, California, who heads the FDA advisory committee on gene therapy, says that scientists must investigate whether such methods are fully effective. But if they are shown to work well, suicide genes might become a standard safety measure in future gene-therapy trials. Observes Salomon: "I think Alain Fischer would be very happy right now if he had a suicide gene in his original vector."

Other researchers are investigating 'insulating' sequences to minimize the interaction between vector and host DNA. The initial idea was to prevent host genes from interfering with expression of the corrective gene. For instance, David Emery at the University of Washington in Seattle has put a genetic sequence from chickens into a retroviral vector, and found that the modification allowed cells that integrated the vector to express more of the therapeutic gene⁷. Hopefully, this buffering effect will also work in reverse, preventing the corrective gene from interfering with host genes.

While they work on new strategies, gene therapists will have to stick with some imperfect techniques. The French case is an individual tragedy, but experts point out that such events are, unfortunately, part and parcel of the business of developing new therapies. "It's no surprise that the more effective the gene-therapy transfer becomes, the more we're going to see the adverse consequences that are intrinsic to that technology," says Freidmann.

Rather than backing away from gene therapy, argue its practitioners, the important lessons to be learned from the French case lie in a deeper understanding of the risks that will help doctors to predict and respond to similar events in the future. ■

Erika Check is Nature's Washington biomedical correspondent.

1. Cavazzana-Calvo, M. *et al. Science* **288**, 669–672 (2000).
2. Check, E. *Nature* **419**, 545–546 (2002).
3. Aiuti, A. *et al. Nature Med.* **8**, 423–425 (2002).
4. Li, Z. *et al. Science* **296**, 497 (2002).
5. Schmidt, M. *et al. Blood* **110**, 2737–2743 (2002).
6. Schröder, A. R. W. *et al. Cell* **110**, 521–529 (2002).
7. Emery, D. W., Yannaki, E., Tubb, J. & Stamatoyannopoulos, G. *Proc. Natl Acad. Sci. USA* **97**, 9150–9155 (2000).

A lens less ordinary

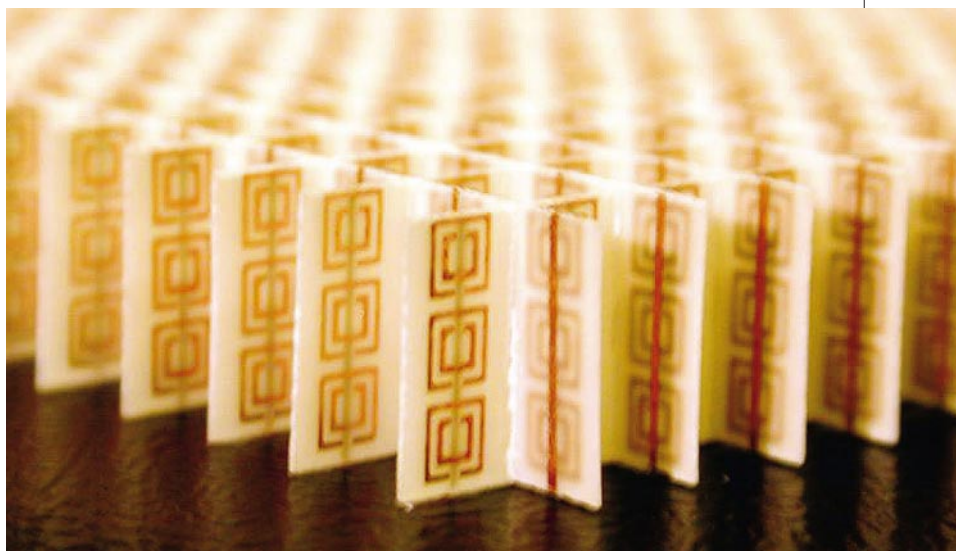
In the 1960s, a Russian physicist considered the properties of a material that didn't yet exist. Now researchers appear to have fulfilled his predictions — but is everything as it seems? Liesbeth Venema investigates.

There are some truths in physics on which we have come to depend. Light rays, for example, bend when they cross the boundary between two materials. That's why an oar dipped into water appears to bend towards the surface, and why the pool itself looks shallower than it really is.

But this familiar phenomenon, called refraction, is beginning to look less straightforward. In the lab of David Smith, a physicist at the University of California, San Diego, a strange array of metal wires and loops has been pieced together. In April 2001, Smith and his team showed that this construction, which they refer to as a 'metamaterial', has a peculiar property: it bends electromagnetic waves in the opposite direction to normal materials¹.

If a pool of water had this property, known as negative refraction, oars would bend away from the surface, and the pool would appear deeper than it really was. But Smith's construction is more than an interesting oddity. Some in the field believe that it could lead to what John Pendry, a physicist at Imperial College in London, has dubbed "the perfect lens" — a device that can produce flawless visual images. "These materials have electromagnetic properties never seen before," says Sheldon Schultz, a colleague of Smith's. "This is a whole new ball game."

The idea of negative refraction first arose during the 1960s, when Victor Veselago, a physicist then at the P. N. Lebedev Physical Institute in Moscow, considered the optical properties of an imaginary material. Every material has a refractive index, which measures how fast it transmits light and how light is bent on entering the material — the higher the index, the slower the propagation and the stronger the deflection. All naturally occurring materials have a positive index — air is 1, glass about 1.5 — but Veselago, who now



Positive result: this experimental 'metamaterial' appears to have a negative refractive index.

heads the magnetic materials laboratory at the General Physics Institute in Moscow, was interested in how light would behave in a hypothetical material with a negative index.

Negative thoughts

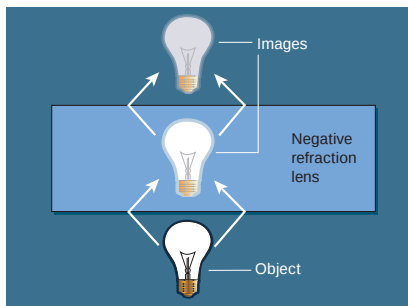
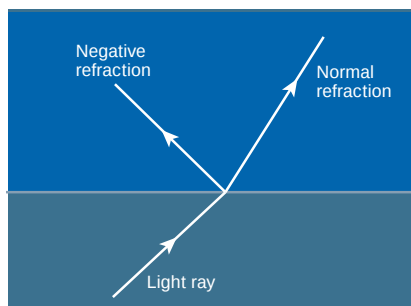
A material's refractive index depends on its response to the electric and magnetic components of an electromagnetic wave, measured by its permittivity and permeability, respectively. Most materials have positive permittivities: place one in an electric field, and the direction of the field induced inside the material will have the same orientation as the applied field. Most materials also have positive permeabilities, and react to magnetic fields in a similar way.

But Veselago was interested in a hypothetical material with a negative permittivity and

permeability, and hence a negative refractive index. The fields inside such a substance would be orientated in the opposite direction to the applied fields. And, as Veselago showed, this has some interesting consequences². Light entering such a material would take a sharp turn at the boundary, for example, so a rectangular slab of material would act as a lens, creating an image inside the slab and then again on the other side of the slab (see diagram).

With no materials available on which to test Veselago's theories, researchers could not follow up his ideas. But during the late 1990s, Pendry and his colleagues at Imperial College, together with researchers at Marconi, a telecommunications and computing company also based in London, began to produce structures with the right kind of properties.

The negative permittivity to visible light of some metals, such as silver, had been established well before Veselago's original studies. Pendry, who was developing devices to control the microwaves used in radar systems, was interested in developing materials with negative permeability. Both permittivity and permeability depend on the collective response of the electrons within a material to the applied electric and magnetic fields. To control this response, Pendry created an array of closely spaced, thin, conducting elements, such as metal hoops, which as a whole behaved as a kind of 'composite' material. In 1999, Pendry described how he adjusted the



Out of line: negative refraction (left) bends light in an odd way, and could be used to create a lens (right).

array's properties, such as the spacing between the elements, to create an array with negative permeability³.

Following up on this work, Smith set about making a material with a negative refractive index. His group created interlocking units of thin fibreglass sheets imprinted with copper rings and wires (see previous page). Last spring, his team sent a beam of microwaves through a small sample of the metamaterial and found that Veselago's original calculation was correct: the microwaves were bent in the opposite direction from normal¹.

Akhlesh Lakhtakia, a physicist at Pennsylvania State University in University Park who had been following the developments, says that he really began to take notice when Smith's results came out. He points out that the microwave beam did not make a small deviation — it turned through almost 80°, which would give the material a refractive index of about -2.7. "This is very convincing," he says.

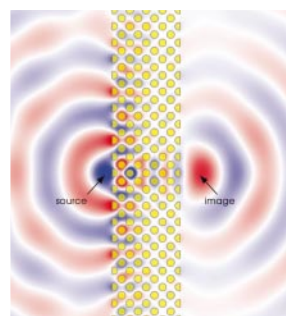
But not everyone agrees. This April, Prashant Valanju and his colleagues at the University of Texas at Austin, who are also working on metamaterials, suggested that there is something fundamentally wrong with the literature on negative refraction⁴. Real electromagnetic waves are made up of a mixture of individual waves of different frequencies, and each component has its own velocity. When researchers measure the direction and strength of the main wave, they actually record the combination of these components.

Valanju's group modelled what would happen when a wave made up of two frequencies enters a material with a negative refractive index. They found that although the components would be refracted negatively, they would combine to give a wave that was positively refracted. The result seems paradoxical, but for physicists who are used to the odd ways in which waves of different frequencies can combine, it is plausible. Valanju's analysis also showed that the different components should quickly spread out and weaken the signal.

According to Valanju's team, Smith's results were an experimental artefact resulting from the fact that he used a thin sample —



Focused: John Pendry's simulations (right) suggest that photonic crystals may have negative refractive indices for visible light.



the phenomenon would disappear if a thicker piece of the metamaterial was used. Lakhtakia agrees that some of the points made by Valanju raise questions about the experiments. When Smith measured the angle of transmission, for example, he placed his detector very close to the sample, and so could have missed the dispersion predicted by Valanju's model.

Frontal assault

But Smith and Pendry responded with their own theoretical analysis of a two-frequency wave. Rather than considering an infinitely wide wavefront, as Valanju had done, they looked at a more realistic, finite beam of radiation, and found that it should indeed be negatively refracted⁵. Their findings are also backed up by recent experiments. Claudio Parazzoli and colleagues at Phantom Works, the Seattle-based research and development unit of aerospace firm Boeing, have built a larger metamaterial. In unpublished work, they describe negative refraction of microwaves, which they managed to detect tens of centimetres away from their metamaterial. Together with Smith and Pendry's study, this is enough to convince most researchers that Valanju's criticisms have been answered and that negative refraction exists.

So if the phenomenon is real, what can it be used for? Pendry began thinking about applications before Smith's study. In 2000, he considered the properties of a rectangular lens made out of a negatively refracting material. Some of the waves emitted or reflected by ordinary objects decay very quickly, preventing normal lenses from transmitting them. As a result, the detail contained by these

'evanescent' waves is lost — even when the highest-quality lens is used. But Pendry predicted that materials with negative refractive indices would amplify evanescent waves, thus retaining the information they contain⁶.

Pendry's idea has since attracted its fair share of comments and criticisms in journals and preprint servers, in part fuelled by his provocative use of the term "perfect lens" to describe the concept. "Maybe the title was not such a good idea," admits Will Stewart, chief scientist at Marconi, who has worked with Pendry. Some researchers say that even a tiny amount of absorption in the lens would prevent amplification of the evanescent waves. Others disagree, and the jury is still out on whether or not a perfect lens can be created.

As the debate rumbles on, other groups are trying to create a material with a negative refractive index for visible light. Interest centres around photonic crystals, materials with alternating regions of different refractive indices. A slab of transparent material with an array of holes drilled into it is one example. Like metamaterials, photonic crystals can be designed to transmit light in different ways, and theoretical studies suggest that negative refraction should be possible^{7,8}.

John Joannopoulos, a photonic-crystal expert at the Massachusetts Institute of Technology, predicts that experimental confirmation will come in the next two years.

Few concrete suggestions for applications have emerged so far, although researchers have noted potentially useful properties of flat lenses. They would be easier to make than conventional curved lenses, for example. They would also create an image of an object placed anywhere along their length, unlike curved lenses, which only focus objects placed in front of them.

For the normally quiet world of electromagnetic materials, such excitement is unusual. But Smith's experimental work, together with Pendry's provocative ideas, has turned a lot of heads. And if one of the groups trying to make a perfect lens succeeds, interest is likely to snowball. Veselago, who continues to work on the theory of magnetic materials, has had to wait many years for his idea to take off, but now it seems to have a life of its own. ■

Liesbeth Venema is a senior physics editor at Nature.

- Shelby, R. A., Smith, D. R. & Schultz, S. *Science* **292**, 77–79 (2001).
- Veselago, V. G. *Sov. Phys. Usp.* **10**, 509–514 (1968).
- Pendry, J. B., Holden, A. J., Robbins, D. J. & Stewart, W. J. *IEEE Trans. Microwave Theory Tech.* **47**, 2075–2084 (1999).
- Valanju, P. M., Walser, R. M. & Valanju, A. P. *Phys. Rev. Lett.* **88**, 187401 (2002).
- Smith, D. R., Schurig, D. & Pendry, J. B. *Appl. Phys. Lett.* **81**, 2713–2715 (2002).
- Pendry, J. B. *Phys. Rev. Lett.* **85**, 3966–3969 (2000).
- Luo, C., Johnson, S. G., Joannopoulos, J. D. & Pendry, J. B. *Phys. Rev. B* **65**, 201104 (2002).
- Notomi, M. *Opt. Quantum Electron.* **34**, 133–143 (2002).

P. VALANJU/UNIV. TEXAS AUSTIN



Prashant Valanju (left) and his team dispute the fact that negative refraction has been achieved.

P. PENDRY

REF. 7

The next challenge is to map the human mind

An ambitious project aims to chart the territory of ideas: vast but, conceivably, not infinite.

Sir—Can we comprehend the ideas and thoughts of other beings? The scientific literature from a variety of fields, including psychology, sociology and ethics, suggests that we can. The methodology used in the disciplines of genetics, psychology, animal behaviour, sociology, history, public understanding of science, and religious studies—to mention only a few—can and should be collected to design an integrative approach to understand the extent of human ideas. Readers are invited to a meeting in Japan on 15 February 2003 to start such a project (see www.biol.tsukuba.ac.jp/~macer/index.html).

If we define an 'idea' as the mental conceptualization of something—including physical objects, an action or sensory experience—then the number of objects in the universe of a living being is finite. Both the number of possible choices for action and the sensory states of animals are finite. In that sense we can expect to be able to count ideas. The initial methodology could separate classes of

ideas, as follows: (1) conceptualization of physical objects; (2) psychological meanings of images associated with objects (such as colours); (3) memories; (4) plans for short- and long-term future; (5) intention to modify one's behaviour; (6) intention to modify behaviour of surrounding beings and the environment; (7) sensory states such as pain, pleasure, libido; (8) inhibition of a response based on immediate evolutionary benefit, for example, memes; and (9) interactive conceptualization of ideas in a community-based response.

I propose a 'mental mapping project' to explore similarities between cultures and communities, both at the individual human level and as members responding inside biological communities. There are implications for cultural identity. How should the culture that tries to maintain its uniqueness face up to the reality that the full range of idea diversity is found in every culture? Idea diversity is found in almost all cultural groups, excluding those formed to promote particular political aims, such as

those who fight for or against the right to choose abortion or euthanasia.

The universality of ideas is important for the development of global society, when we're faced with dilemmas such as whether to have common international guidelines to regulate the use of new biotechnology, for example, or of assisted reproductive technology using cloning. It is time to start thinking scientifically about this in a coordinated way.

Although the human mind appears to be infinitely complex, and the diversity of humankind and culture has been considered vast, I would suggest that the number of ideas that human beings have is finite. Hence my call for a project to map the ideas of the human mind. We already have the means to embark upon a human mental map with the goal of describing the diversity of ideas a human being makes in any given situation or dilemma.

Darryl R. J. Macer

Institute of Biological Sciences, University of Tsukuba, Tsukuba Science City 305-8572, Japan

Moving beyond 'industry vs ecologists' stereotype

Sir—I agree with your statement in your News story about my 'superweed' research (*Nature* **419**, 655; 2002) that academic researchers need better access to pre-commercial transgenes to be able to carry out independent, empirical research on environmental risks of genetically modified organisms (GMOs). Yet the fragile relationship between seed companies and university researchers is improving, and readers may not get this impression from your story.

A small community of ecologists and population geneticists is working with agricultural biotechnology companies and regulatory agencies to assess the environmental effects of transgenic crops. In my case, two multinational companies invited my colleagues and me to evaluate the effects of pre-commercial *Bt* sunflowers on wild sunflowers, which are a common weed in the midwestern United States. The companies agreed to fund a research project that would be published independently in peer-reviewed journals (A. A. Snow *et al.* *Ecological Applications* (in the press)). They provided us with a *Bt* transgene in a cultivated sunflower, as well as valuable technical assistance.

This liaison was not always easy, as companies are bound by confidential business plans whereas academic

ecologists are eager to talk about work in progress. I'm glad we carried out this research, even though it was frustrating when we were not allowed to continue using university funds. The companies had decided not to commercialize this variety of *Bt* sunflower, but we wanted to carry out a larger-scale project because so little is known about the ecological and evolutionary effects of transgenes that could spread throughout natural populations.

Unfortunately, polarization of views on GMOs often hinders the use of ecological research in risk assessments. Seed companies have been slow to acknowledge that ecological studies are needed to evaluate some of their products. They also fear the media's tendency to emphasize unwelcome findings and ignore results showing environmental benefits or 'no impact'. At the other end of the spectrum, objections from advocacy groups have reached the point where risk assessment research is often blocked by regulatory delays or eco-terrorism, especially in Europe.

Much of this debate hinges on ethical and political concerns, which are outside the realm of ecological science. Nonetheless, academic ecologists can help to answer questions that arise about the environmental effects of GMOs. I urge that we progress beyond the standard stereotyping of 'industry versus ecologists' to facilitate more effective dialogue and to

introduce sounder science to the debate.

Allison A. Snow

Department of Evolution, Ecology and Organismal Biology, Ohio State University, Columbus, Ohio 43210, USA

Liberal world of science

Sir—Raymond Pierotti's Correspondence expressing his opinion (*Nature* **419**, 667; 2002) that academic scandals are less well known than business ones reminded me of Michael Faraday's description of his "desire to escape from trade, which I thought vicious and selfish, and to enter the service of Science, which I imagined made its pursuers amiable and liberal".

Faraday, a bookseller's apprentice, had written to Sir Humphry Davy at the Royal Institution of Great Britain in pursuit of a career move. Davy advised him that experience would correct his opinion of the "superior moral feelings of philosophic men" (see *The Philosopher's Tree*, ed. P. Day, 2–3; IoP, Bristol, 1999).

Business and science have long been interdependent. Faraday's "philosophic" researches, for example, paved the way for vast technologically based industries and many other scientific, technological and business developments.

Geraint Day

Policy Unit, Institute of Directors, 116 Pall Mall, London SW1Y 5ED, UK

Living with capitalism

Similar processes affect cells and businesses. But what price sustainability?

The Hidden Connections: Integrating the Biological, Cognitive, and Social Dimensions of Life into a Science of Sustainability/The Hidden Connections: A Science for Sustainable Living

by Fritjof Capra

Doubleday/HarperCollins: 2002. 288 pp.

\$24.95, £20

Crispin Tickell

Not for the first time, Fritjof Capra makes a broad sweep of human knowledge and experience, ranging here from a definition of life (or rather the “defining characteristics of living systems”) to global capitalism as we now know it. It is an exciting journey by way of biology, sociology, market economics and ecology, and arrives at conclusions that call for radical change in human society.

He draws throughout on certain principles that are as applicable to the living cells of the smallest organisms as they are to business corporations and political structures. This indeed is consilience on the grandest scale. At its root are ideas about self-generating networks among elements in a system, nonlinear evolution, which often produces the unexpected, and the emergence of new forms of order out of apparent instability.

From all this, Capra develops the idea that “the interactions of a living system with its environment are cognitive interactions, and the process of living itself is a process of cognition”. Here he uses cognition in the sense of the ability to react to perturbations in the environment. Thus “mind and matter no longer appear to belong to two separate categories but can be seen as representing two complementary aspects of the phenomenon of life... At all levels of life, beginning with the simplest cell, mind and matter, process and structure, are inseparably connected.” He goes on to show that throughout the history of life, the planetary web has expanded through mutation, exchange of genes and symbiosis, producing forms of life of ever increasing complexity and diversity. This embraces elements of Gaia theory.

So much for the biology. Its application to human society is inevitably incomplete and somewhat awkward, but networking, nonlinear change, complexity and the emergence of specific cultural identities are evident wherever we look. The same can be said to some degree about commercial organizations and their management. But here the all-too-human search for top-down design can militate against natural forces. Managers like algorithms for success even if they do not



Global village: the spread of capitalism has led China to join the World Trade Organization (WTO).

exist. For many people, machines are a better model than living organisms. No wonder then that so many modern organizations, and even capitalism itself, are under strain.

In the second part of his book, Capra analyses the many shortcomings of global capitalism, using biotechnology, in particular the development of genetically modified organisms, as an example. He argues that globalization in its present form is the result of the computer revolution and the introduction of information technology into almost all aspects of human affairs. This is machine philosophy and technology in the extreme. It occurs with the increasing exploitation of the Earth's resources and corresponding damage to the environment, and can be seen most clearly in the shuttling back and forth of money worldwide at unimaginable speed. It is beyond the control of national governments, whose own power is declining, and even of most corporations and financial institutions. The result is not the sort of evolution associated with living organisms, but rather social alienation, increasing division between rich and poor, and cumulative damage to human values and cultural diversity.

What, then, is the answer? Inevitably, this is the weakest part of the book. It has long been obvious that the kind of society that has developed with industrialization is unsus-

tainable in its present form, and cannot be extended as it is to the rest of the world. An eloquent statement on this point was made in the Amsterdam Declaration on Global Change, published after a conference attended by over 1,000 scientists from the great global research programmes in July 2001. In short, we know what most of the problems are, and we probably know most of the answers, including the application of different technologies and the creation of new and more appropriate international institutions. The difficulty is how to get from here to there. Radical change may already be on its way, but it remains mostly on the fringes. We still measure things wrongly, and here economists have a big responsibility. Unfortunately we may need a catastrophe or two to bring about the fundamental changes that are required.

This book is not easy reading. Sometimes the case is obscured by jargon and unnecessary academic defensiveness. Many of the ideas could have been more simply and economically expressed. But it is a rich resource that should be widely drawn upon. By establishing the connections in its title, which are indeed too often hidden, the author has courageously put together a real tract for our times.

Crispin Tickell is at Abington Old Barn, Abington, Cirencester GL7 5NU, UK.

AP PHOTO/STR

A political view of ecology

Imperial Ecology: Environmental Order in the British Empire, 1895–1945

by Peder Anker

Harvard University Press: 2002. 352 pp.

\$59.95, £41.50

Michael Ruse & Nils Chr. Stenseth

Until about 50 years ago, those writing about the history of science tended to be distinguished elderly scientists who had done their creative work and in retirement were writing about their heroes — and themselves. There was a tendency to see history as a progress to the truth, and the emphasis was on ideas. In tune with the positivist philosophy of the age, science was taken to be a struggle to wrest the disinterested truth from objective nature.

All of that changed when the history of science became a profession. Students were trained as real historians, sensitive to the values of the past as worthwhile in their own right. Archives were opened showing that the private face of science was often a lot less polished and savoury than the public one. And because practitioners now often had little direct experience of science in the laboratory or in the field, there was a consequent de-emphasis of ideas for their own sake.

Imperial Ecology, which started life as Peder Anker's PhD thesis in the History of Science Department at Harvard University, is very much in the new style. The story is about ecology — the study of the relationships between organisms — and its political and cultural history, particularly as practised in South Africa during the first half of the twentieth century. This is a story about people with social agendas and philosophies that they read into their science and then promptly read right back out again. Essentially, the book describes how ecological thinking might influence political views — and how political views may influence ecological thinking.

Anker sees two main groups or ideologies, the people and the ideas being virtually one and the same. On one hand, there were those of the Englishman Arthur George Tansley and many of the pioneers of modern ecology, such as Charles S. Elton, who subscribed to some form of mechanistic philosophy. These people wanted to understand the dynamics (and persistence) of ecological systems, to achieve control over nature. The basic idea seems to be one of 'equilibrium': as things get thrown out of balance by external disruptions (of natural or human origin) and then strive to regain balance and order.

Opposed to this view were the group surrounding the South African statesman and philosopher General Jan Smuts, whose inter-



World view: Jan Smuts' ecological outlook was heavily influenced by his political position.

ests in ecology were primarily related to their political world views. Although he was prime minister of his country for many years (and an active politician for almost all of his life), Smuts was always interested in the underlying metaphysical principles of reality, and to this end he formulated his philosophy of 'holism'.

Holism bore many similarities to the vitalistic philosophy of Henri Bergson and the organismic one of Samuel Alexander and Alfred North Whitehead. Smuts saw the whole of nature as being integrated, a living whole in which humans have a central (but by no means exclusive) role, evolving up from the primitive to the complex. This was bound up with his thinking about black people, whom he considered to be lower down the scale than whites; although he thought them deserving of some respect and opportunities (in this, Smuts was a lot more liberal than many of his countrymen), he certainly did not think them deserving of the full position of those of European descent. Extending this to organic nature, Smuts and his followers stressed its oneness with everything, and how things are naturally in harmony and only at war when there are unnatural disruptions. In a way, darwinism was turned on its head, for the natural state is for things to be interconnected and working together for the whole.

Anker takes these two positions and traces them to the end of the Second World War, when Smuts was at his height as an

international figure for his work in drafting the charter for the United Nations. Holism fell on hard times, as the Nationalists gained power in South Africa and the nation turned towards Apartheid and the evils thereof. But within the field of ecology, holism is still alive — and still troublesome.

There is certainly much of interest in this book, and it is written with great attention to detail and to important concepts and ideas. But at the end one is left wondering about the book and the science that it is discussing. Is ecology really little more than an epiphenomenon of previously held philosophical beliefs? Or does it have a life of its own? Does it have any relationship to sciences that have real experiments, measurements, observations and tests, and revisions in the light of the evidence? Was ecology then, or indeed now, little more than an excuse for people to sound off about their views on the meaning of life (like today's more political ecology movement and ecophilosophy)? Or is there something more, something altogether sterner and more empirical — something that is masked and ignored by the approach taken by the new historians of science? Would the older, less sophisticated historians have found something that is now missed? Could they have been right in their practice, if wrong in their philosophy, showing that not all change is indeed progress?

We would have liked the book much

more if it had included more thorough presentations and discussions of the parallel development of ecology. Surely the philosophy of Smuts lives on in ecology today (as holistic or system-orientated ecology), especially among those who share his vision of an integrated organic world in which there is a natural harmony disturbed primarily by humans. But the mechanistic equilibrium position corresponding to the development of population and community intrinsic feedback interactions is currently the dominating one — not least because it pays proper attention to darwinian evolution.

Perhaps if this book had been written by practising ecologists with proper training in the history of science, or by a historian trained in ecology, it might have covered the ideological and subject-based conceptual issues in a more balanced way. ■

Nils Chr. Stenseth is in the Department of Biology, University of Oslo, N-0316 Oslo, Norway; Michael Ruse is in the Department of Philosophy, Florida State University, Tallahassee, Florida 32306, USA.

Bringing speciation to the surface

Adaptive Radiation of Blind Subterranean Mole Rats

by E. Nevo, E. Ivanitskaya & A. Beiles Backhuys: 2001. 200 pp. 37 euros

Ernst Mayr

The blind subterranean mole rat of the Near East, around the eastern Mediterranean, is an extraordinary taxonomic phenomenon. It was first thought to be a geographically variable single species (*Spalax ehrenbergi*) with local subspecies and karyotypes. But then it was discovered that this set of populations is a compound 'superspecies' of 12 allospecies, each occupying a geographically distinct region, which makes it unique among terrestrial vertebrates. When two of these allospecies meet, they interbreed to form a narrow hybrid belt. But the F_1 hybrids do not interbreed with the parent allospecies (backcrossing). The range of *S. ehrenbergi* (*sensu lato*) is therefore a mosaic of 12 allospecies.

Eviatar Nevo and his colleagues at the Institute of Evolution at the University of Haifa have studied this remarkable pattern for nearly 40 years. Four allospecies occur in Israel, distinguished by increasing numbers of chromosomes ($2n = 52, 54, 58$ and 60) in the increasingly arid climate towards the east and south. Each allospecies is adapted to its home area, as established by painstaking physiological analysis. This work led to the publication of a book, *Mosaic Evolution of Subterranean Mammals* (Oxford University Press, 1999).



Underground evolution: the blind subterranean mole rat is remarkably well adapted to its habitat.

The same team has now published another book that deals with the origins of the unique population structure of *S. ehrenbergi*. Nevo and colleagues have focused on two questions: what peculiar process of speciation led to the extraordinary mosaic pattern of distribution of the 12 allospecies? And why is the adaptation to the gradually changing climate not gradual but stepwise, along the lines of changing chromosome number?

The answer to the first question is that speciation takes place by peripatric speciation in small, peripherally isolated founder populations — the founder effect means that, because the population represents only a small part of the total gene pool, it soon becomes genetically distinct. It can establish contact with the parent species only after the isolating mechanisms have been perfected.

The second question has a more remarkable answer. It turns out that each allospecies is adapted not only to the subterranean mode of life, but also to the particular external environment (climate) where it lives. This adaptation is in part gradual, like the climatic change, but also occurs in steps marked by chromosomal rearrangements, suggesting a remarkable cohesion of the genotype, reorganized by each bout of peripatric chromosomal speciation.

Seminal Nature in Japanese

Chi No Rekishi (Knowledge From Nature)

Just published in Japanese, by Tokuma Shoten, Tokyo, is a collection of *Nature* papers describing 21 significant scientific advances of the twentieth century. Each paper is accompanied by an illustrated essay by a luminary in the field concerned, placing the development in context. The collection is topped and tailed by announcements of the discovery of *Australopithecus africanus* in 1925

Nevo's admirably thorough analyses of numerous populations, both closely and more distantly related, shed light on such controversial subjects as function, origin and adaptationist programmes. This work is a model of modern population analysis. ■

Ernst Mayr is at the Museum of Comparative Zoology, Harvard University, Cambridge, Massachusetts 02138, USA.

Trying to figure it out

The Millennium Problems: The Seven Greatest Unsolved Mathematical Puzzles of Our Time

by Keith Devlin

Basic Books: 2002. 256 pp. \$26, £18.99

Jeremy Gray

The millennium problems of the title are seven mathematical puzzles proposed in 2000 for solution by the Clay Mathematics Institute in Cambridge, Massachusetts. They are drawn from different areas of mathematics, and each carries a prize of US\$1 million for whoever solves it. While some might wish to see other problems in their place, or at least funded to the same extent, no one would dispute their importance, or deny that a solution would be a significant advance in mathematics. Such is their difficulty and importance that the widely distributed joke is that most mathematicians would pay a million dollars to be able to solve one of them.

To take them in the order they are given here, the first is the Riemann hypothesis, which concerns the distribution of prime numbers. The second asks for a rigorous mathematical theory of the Yang-Mills equations in particle physics, which are capable of establishing some of the key beliefs of physicists about mass. Problem three is about the complexity of problems solved by computers. Number four is about the mathematical equations governing fluid flow, which are still essentially unsolved.

The fifth problem is about a conjecture

and the cloning of Dolly the sheep in 1997.

In between, among other notable events, are the discoveries of the neutron, nuclear fission, seafloor spreading, pulsars, T-cell restriction, buckminsterfullerene and extrasolar planets. There is also a proposed structure for DNA from 1953 that has stood the test of time.

The book is entitled *Chi No Rekishi (Knowledge From Nature)* and is edited by Laura Garwin and Tim Lincoln. It will appear in English next year.

that certain basic methods of algebraic topology can unfailingly recognize a (three-dimensional) sphere — curiously, the problem has been solved in all other dimensions. The sixth is about a remarkable conjecture in algebraic number theory, not far from the circle of ideas that led Andrew Wiles and Richard Taylor to the solution of Fermat's Last Theorem. The final problem returns to algebraic topology, but with a more technical question about the closeness or otherwise of algebraic and geometric ideas in versions of that theory. These last three are harder to 'sell' to a broad audience, but they each point to something quite specific in a major branch of mathematics which for some profound reason we do not understand.

I believe that mathematics should set out its stall in the way that the Clay Mathematics Institute has done, but it would be a valuable educational task to make mathematics graduates aware of what these problems say. The longer accounts of each problem on the Clay Mathematics Institute website (<http://www.claymath.org/index.htm>) presume a considerable amount of background knowledge.

But the problem of exposition is much more acute when, as Devlin has done, the decision is to aim for the largest audience possible.

Devlin's solution was to write a very elementary chapter on each problem, sometimes with a slightly harder set of short appendices. The result is a series of superficial remarks largely intelligible to schoolchildren over 16, followed by a selection of more insightful remarks, followed by some vigorous hand-waving in the direction of the summit. These last are often of a motivational kind, supported by loosely accurate historical remarks, and aimed at amplifying the claim that the problems matter. The quality of the mathematical exposition is high, a sense of excitement is strongly conveyed, and the summits are at least glimpsed while the difficulties in approaching them remain rightly shrouded in mist. There will be other books, by other authors, aimed at readers who already know more mathematics.

Devlin has grappled openly and honestly with the problems of reaching his chosen audience. He is frank in admitting that it is easier to hint at some of the problems than

others, and this accounts not only for the order in which he presents the problems but the sort of things he says about them. Full marks for trying. I hope, however, that readers of *Nature* will find this book too easy and want to read something deeper. It would be a poor reflection on scientists' education if they did not, and worse if they could not.

But I hope that bright, mathematically inclined school children will read this book and turn to mathematics. They might even go on to solve one of these problems, which would be magnificent. Even if all seven problems are eventually solved, mathematics contains many more problems that are at least as intellectually rewarding. ■

Jeremy Gray is at the Centre for the History of the Mathematical Sciences, Open University, Milton Keynes MK7 6AA, UK.

More on mathematics

Dr Riemann's Zeros: The Search for the \$1 million Solution to the Greatest Problem in Mathematics

by Karl Sabbagh
Atlantic Books, £14.99

Science in culture

A physical response to architecture Two Swiss architects design rooms to change your hormone levels.

Alison Abbott

Architecture is about space and light, usually considered in the abstract. But Swiss architects Philippe Rahm and Jean-Gilles Décosterd are fascinated with how these properties affect us physically. In a series of projects over the past four years, they have examined the relationship between the external environment and human physiology, for example by modifying the composition of the air, or the wavelength of the light, that fills their architectonic space. They are aware that this is bringing architecture to an ethical crossroads.

This year they were selected to represent Switzerland at the Eighth International Architecture Exhibition at the La Biennale de Venezia, which closed recently, for which they designed their 'Hormonarium'. Fittingly, the Hormonarium transports a high-Alpine environment to sea level, in much the same way that an urban swimming pool might be considered to be bringing the environment of a lake into a city.

The Hormonarium is a plain white room containing four white sofas. The visitor enters through a system of air-tight double doors. The floor of the room is imbedded with long rows of bright, full-spectrum fluorescent lights. Lit from below, the extreme brightness of the room evokes the snow-reflected light on a mountain-top on a sunny day. The ultraviolet part of the spectrum tans the skin, the infrared part could burn it. The

level of oxygen in the air is reduced from the normal 21% at sea level to just 14.5%, as it would 3,000 metres up a mountain. After 10 minutes, the visitor experiences the pleasant light-headedness that accompanies a successful climb.

Playing further with the invisible parameters of the environment, the architects fill the room with 'music' of such low frequency (40–80 Hz) that it is felt through vibration, rather than heard. Visitors must wear blue plastic overshoes that crunch on the glass floor, adding to the sensation of being surrounded by snow.

The Hormonarium is so-called because the low-oxygen environment raises levels of erythropoietin, a hormone that stimulates the production of red blood cells, at least after a few hours. This is the basis of high-altitude training, where athletes train in the mountains to pump up the concentration of their red blood cells, so they can use oxygen more efficiently during a race.

This is not the architects' first engagement with hormone manipulation. Their 'Melatonin Room', exhibited last year at the San Francisco Museum of Modern Art, was alternately filled with an intense green light and a weak ultraviolet light, which suppress the production of the sleep hormone melatonin to different degrees. But if all this sounds like installation art rather than architecture, Rahm would disagree. He considers himself an architect first and foremost: "I want to work in the real world," he says.

Together with Décosterd, Rahm is designing a 'winter house', commissioned by a French artist and to be built near France's Atlantic coast. The house recreates a tropical environment in a very



Mountain high: the Hormonarium's Alpine oxygen level boosts erythropoietin production.

literal way. The lighting will mimic, in intensity and timing, the natural day-and-night cycle of French Polynesia. The air will be as warm and moist as in Tahiti — and also just as scented, because the room that heats the house's circulating air will be packed with Tahitian plants. With its sealed and dislocated environment, the house will be used for just a few winter days at a time.

Rahm and Décosterd are moving ahead faster on the philosophical, rather than the practical, front with their physiological architecture. "There are ethical barriers to the wide use of architecture to modulate mood," says Rahm. ■

Alison Abbott is Nature's senior European correspondent.

JEAN-MICHEL LANDECY

Back to the beginning

Wolf Reik and Wendy Dean

A newly fertilized egg has the ability to differentiate into all tissues of the embryo, a state known as totipotency. A process known as epigenetic reprogramming returns the highly differentiated sperm and egg nuclei to this nascent state in the early embryo. Somatic nuclei that are introduced into eggs in order to clone animals must also undergo epigenetic reprogramming, and recent work shows that the failure to do so may be the chief obstacle to cloning success.

The epigenetic view of development has long surpassed the theory of preformation (in which the gametes contain small but perfectly formed bodies waiting to grow). Epigenesis (or, in modern terms, lineage commitment) holds that differences in cells and tissues arise in development because gene-expression programmes change as cells differentiate. This view of development was most famously expressed by C. H. Waddington, who likened the path of a cell lineage towards terminal differentiation to a ball travelling downwards along branching valleys; once it has entered its final valley it cannot easily cross the mountain into the neighbouring one (transdifferentiation or plasticity) or return to the beginning (cloning or return to totipotency).

However, the differences in gene expression occur without any change in the sequence of DNA, which would render the process completely irreversible. Thus, although the model clearly envisaged difficulties with transdifferentiation or cloning, the reasons for these difficulties remained uncertain. The acid test of the prediction that changes in gene expression take place without underlying changes in the DNA sequence was to attempt to clone an animal by taking a differentiated nucleus (from the intestinal epithelium, say) and introducing it into an egg

from which the nucleus had been removed. If the specialized transcriptional programme in a somatic nucleus is determined by the soup of transcription factors available to it, then by exposing the nucleus to the different soup that is present in the egg, reprogramming can occur. This was indeed demonstrated in the pioneering experiments by J. B. Gurdon, in which frogs were cloned from specialized cells, and more recently in several mammalian species (including Dolly the sheep). But despite these successes, the efficiency of cloning remains very low in all species in which it has been attempted. Has the right reprogramming soup not yet been found?

As well as changes in the availability of transcription factors and in gene expression, the genome undergoes physical epigenetic changes (the central topic of the science of epigenetics). Methyl groups are attached to C–G dinucleotide sites in the DNA, and the histone proteins around which the DNA is wrapped undergo chemical modifications such as acetylation, phosphorylation, methylation and ubiquitination. Specific combinations of these histone modifications can mark genes for activity or silencing, and histone methylation can combine with DNA methylation to reinforce its repressive effect on gene activity. The key property of these epigenetic marking systems is that they are thought to retain stability as cells divide. Thus cells rolling down Waddington's differentiation valleys accumulate epigenetic markings that keep some genes active and others silent. The problem of reprogramming has therefore become a matter of how to reset these markings.

The best place to look for reprogramming is where it takes place naturally, in the formation of gametes and in the zygote. There must be an epigenetic reprogramming mechanism by which cells in some of Waddington's valleys (those that result in sperm and eggs) can return to the beginning.

There is some evidence of epigenetic reprogramming in action in early embryos and in gametes. Sperm DNA experiences a drastic loss of DNA methylation only hours after fertilizing an egg, and before its DNA is replicated, but the egg-derived nucleus remains highly methylated. Further demethylation and reorganization of histone modifications takes place as cells in the early embryo divide, before the first differentiation steps begin. After that, and concomitant with differentiation, high levels and cell-type-specific patterns of DNA methylation and histone modifications are acquired. So do these create an epigenetic momentum that is difficult to reverse during cloning?

There is evidence that epigenetic reprogramming is aberrant in clones. Methylation

Epigenetic reprogramming

Efficient cloning will require the ability to reset the gene-expression programmes of specialized cells, in the same way that sperm and eggs form undifferentiated embryonic cells.

levels remain high in the majority of early cloned embryos. Morphologically, the nuclei in cloned embryos resemble those of the somatic cells that were used for cloning, which probably indicates that epigenetic modifications have not been removed properly. The early signs are that epigenetic reprogramming is inefficient, which may explain many of the problems encountered by cloning attempts.

The idea of epigenetic reprogramming opens up exciting possibilities for the development of new medical avenues, such as therapeutic cloning and stem-cell therapies. First, it is important to determine whether monitoring reprogramming in individual cloned embryos will provide a diagnostic tool to select those embryos with the best chance of growing into a healthy organism. Second, the study of the mechanisms of reprogramming as it naturally occurs in the early embryo (particularly the processes of demethylation and remethylation) should provide a means of intervention in either the cells used for cloning, or in the early cloned embryos. Finally, the possibility of moving cells from one of Waddington's valleys into another (transdifferentiation) has recently received much attention, with several studies suggesting that cells are more plastic and pliable than we thought.

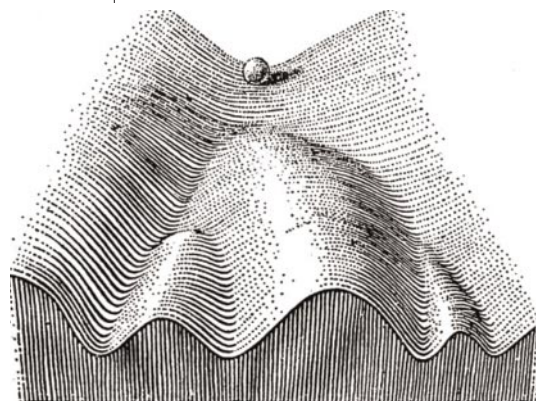
Transdifferentiation will also involve epigenetic reprogramming, although the mechanisms and changes in patterns may be less dramatic than those involved in returning to totipotency. Understanding how Waddington's cells travel down the valleys of differentiation, and how they can be returned to their starting point both naturally and by experiment, will help us to unlock many exciting possibilities in biology and medicine. ■

Wolf Reik and Wendy Dean are in the Developmental Genetics Programme, The Babraham Institute, Cambridge CB2 4AT, UK.

FURTHER READING

Waddington, C. H. *Organisers and Genes* (Cambridge Univ. Press, Cambridge, 1940).
Gurdon, J. B. & Uehlinger, V. *Nature* **210**, 1240–1241 (1966).
Solter, D. *Nature Rev. Genet.* **1**, 199–207 (2000).
Reik, W., Dean, W. & Walter, J. *Science* **293**, 1089–1093 (2001).
Blau, H. M. *Nature* **419**, 437 (2002).

WADDINGTON, C. H. THE STRATEGY OF THE GENES (ALLEN & UNWIN, 1957)



On a roll: the process of epigenesis has been likened to a ball falling into one of several valleys.

Metabolic balance sheets

Athel Cornish-Bowden and María Luz Cárdenas

Application of book-keeping principles to metabolic networks provides a powerful technique for understanding the properties of microorganisms and predicting the results of genetic modification.

Stoichiometric analysis has the same function in biochemistry as book-keeping has in business. Both deal with the inputs and outputs in flow systems, and studying the stoichiometric structure of a metabolic network might seem only slightly more exciting than a visit to an accountant's office. Times change, however, and progress in systems biology will involve paying closer attention to stoichiometry than seemed necessary in the past. The benefits should be that, with minimal knowledge of kinetic parameters, we will be able to predict how systems will respond to changes in conditions, and how they can be genetically engineered to produce desirable characteristics.

The growing knowledge of genomes of different organisms has brought new life to the study of metabolic networks, and a striking example appears on page 190 of this issue¹. Stelling and collaborators discuss there the idea that, by breaking a network down into 'elementary flux modes', based on simple accounting for metabolic inputs and outputs, possible properties of the network can be predicted.

A typical biological network such as the central metabolism of the gut bacterium *Escherichia coli* consists of many processes that operate simultaneously and in parallel. For example, many different metabolic products are being synthesized at different rates at the same time as substrates such as glucose are being consumed to supply power for all the activity. Even when all the individual processes have been identified, it is no trivial matter to predict how the properties of the network as a whole will change if the activity of one enzyme is changed.

It used to be widely assumed, for example, that complete elimination of an enzyme activity would have obvious effects, but this expectation has been overturned by observations of the effects of gene knockouts in various organisms: in *E. coli*, fewer than 300 out of 4,000 genes are 'essential' in the sense that deleting one of them prevents growth on a rich medium². Many of the others can be deleted (one at a time) without producing any evident consequences, even for growth on a medium containing restricted nutrients; in other words, they are 'silent'. As well as discovering the functions of these silent genes, it is crucial to know why they are not essential, and elementary flux modes provide a tool for addressing this question.

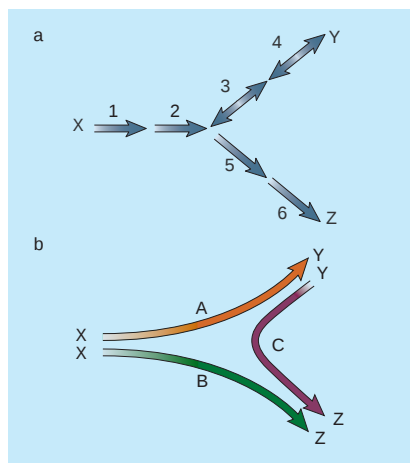


Figure 1 Networks made simple. a, A network of six reactions, of which reactions 3 and 4 are reversible. b, This system can be decomposed into three elementary flux modes, A, B and C. It is evident that transformation of input X into output Y is a function of mode A, and that deleting a reaction that is not part of it (for example number 5) will not prevent this transformation from occurring. This type of analysis has been taken to sophisticated levels by Stelling *et al.*¹, who in their study of *Escherichia coli* metabolism looked at 43,279 elementary flux modes.

Consider, for example, the simple branched network of six reactions shown in Fig. 1a. There are three subnetworks that account for everything the whole network can do. These are the elementary flux modes, labelled A, B and C in Fig. 1b, of which mode A consists of reactions 1, 2, 3 and 4; mode B of reactions 1, 2, 5 and 6; and mode C of reactions 3, 4, 5 and 6. Analysing the network in this way makes it immediately obvious that transformation of X into Y is a function of flux mode A and that deleting a reaction that is not part of it, for example reaction 5, will not prevent the system from achieving this transformation. Predicting whether a mutant will be able to grow involves identifying elementary flux modes that do not use the missing reaction. In this simple example, the analysis merely confirms what is evident from inspection, but in larger systems this is not the case.

Systems of moderate size have previously been decomposed into elementary flux modes^{3,4}. But there has been no analysis of a system that encompasses most of an organ-

ism's metabolic activity, something that is necessary if unintuitive properties are to be deduced. The approach has now come of age with the analysis by Stelling and colleagues¹ of a network of 110 reactions, linking 89 different metabolites, that encapsulates our knowledge of the central metabolism of *E. coli*. This system is much larger than any analysed previously, and its complexity is illustrated by the huge number — 43,279 — of elementary flux modes, far beyond the range of analysis possible by inspection.

The results of Stelling *et al.* allow us to understand some of the experimental properties of *E. coli* in terms of its stoichiometric structure. For example, glucose is involved in 27,099 elementary modes, more than twice as many as glycerol, in agreement with biochemical experience that glucose is the more important energy source. Moreover, around 7% of the elementary modes for glucose allow energy production and growth in the absence of oxygen. But none of the elementary modes for glycerol, succinate or acetate allows this, in agreement with the observation that *E. coli* can grow anaerobically on glucose but not on these other energy sources.

The agreement between theory and experiment provides a basis for believing that this method can produce valid conclusions even without knowing the answer in advance, and so can be used for organisms that are less well known than *E. coli*. Specifically, we can now predict which mutations, in which combinations, a bacterial or yeast culture can tolerate, and what genetic modifications (addition or suppression of catalytic activities) will permit new properties to be created. Moreover, as flux modes may differ in the amount of product they generate per unit of input, they allow insight into how yields may be improved. Organisms such as *E. coli* that can survive in widely varying conditions owe their metabolic robustness, or resistance to mutation, to a high degree of redundancy — in other words, they have a much larger number of elementary flux modes than are absolutely necessary for growth in any particular conditions. More specialized organisms will certainly have less redundancy in their design and fewer elementary flux modes.

Until now, biochemical progress has been driven mainly by new observations rather than theories — in contrast, say, to modern

physics. However, knowledge of elementary flux modes allows predictions of what should happen after a particular genetic manipulation or change in culture conditions, and these predictions can, of course, be tested. Stoichiometric analysis will thus give a boost to hypothesis-driven experimental research in biology. ■

Athel Cornish-Bowden and María Luz Cárdenas are in the Laboratoire Bioénergétique et Ingénierie

des Protéines, Centre National de la Recherche Scientifique, 31 chemin Joseph-Aiguier, BP 71, 13402 Marseille Cedex 20, France.
e-mails: athel@ibsm.cnrs-mrs.fr
cardenas@ibsm.cnrs-mrs.fr

1. Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S. & Gilles, E. D. *Nature* **420**, 190–193 (2002).
2. Cséte, M. E. & Doyle, J. C. *Science* **295**, 1664–1669 (2002).
3. Schuster, S. et al. *Bioinformatics* **18**, 351–361 (2002).
4. Van Dien, S. J. & Lidstrom, M. E. *Biotechnol. Bioeng.* **78**, 296–312 (2002).

Earth science

Is it all in the crust?

Simon Lamb

New observations suggest that earthquakes on land are only 'skin-deep', confined to the Earth's outermost layer of crust. This has prompted a rethink of what gives the tectonic plates their strength.

Writing in *GSA Today*, James Jackson¹ is seeking to overturn more than two decades of geological thinking about the movement of the Earth's surface. He proposes that where these motions occur on land, we need look no deeper than the crust for their driving forces. If he is right, then it is going to be difficult to predict the long-term effect of these forces — and the occurrence of earthquakes.

In the late 1960s, with the advent of the theory of plate tectonics, geophysicists finally seemed to have made sense of earthquakes. The surface of the Earth is covered in a mosaic of rigid plates, and earthquakes, in this theory, occur only in narrow zones where the plates rub against each other. This was good news, because it provided a basic understanding of why there were earthquakes in the first place. But in the 1980s, there were second thoughts.

The plate theory worked well in the oceans, but across the continental land masses, earthquakes seemed to occur all over the place. Gradually a new view emerged that wasn't plate tectonics at all: in fact, large tracts of the continents seemed to behave more like a fluid².

The idea was that the movements along fault lines in the continental crust were driven by a deeper, fluid-like flow in the underlying mantle³ (Fig. 1a). This opened up new possibilities for earthquake prediction, because the well-known physics of fluid flow, in combination with the flow properties of the mantle, could be used to understand those movements. But if mantle rocks do have such a dominant influence, they must be generally stronger or stiffer than those in the crust.

Jackson, who was himself involved in developing this fluid theory of continents, has now changed his mind. He believes that

we must go back to trying to understand the motions of small plate-like blocks in the crust — bad news for those studying the active fault lines, because, as yet, there is no simple way of predicting how these blocks might behave.

To justify such a paradigm shift, good evidence is needed. It has come in the form of new observations about the nature of the Earth's outer shell, comprising the crust and uppermost mantle and called the lithosphere. Jackson's argument rests on the idea that a portion of the continental lithosphere is elastic and can be bent or flexed, supporting substantial stresses, rather like a springy beam. This beam must hold a substantial part of the strength or stiffness of the lithosphere. A critical question is where in the lithosphere this elastic beam lies.

It now seems that all earthquakes in the continents occur only in the crust^{4,5}, not in the mantle. This important discovery has emerged from an increased number of seismic studies, combined with improved computer techniques to model the data. As earthquakes can only occur where the Earth shows springy or elastic behaviour, Jackson argues that the brittle crust is itself the springy beam. And he has more evidence up his sleeve.

McKenzie and Fairhead's study⁶ of the strength of gravity in northern India shows that this region has bent under the weight of the Himalayas as though it is an elastic beam about 40 km thick — a thickness that nearly matches that of the crust (40–45 km). If you bend an elastic beam, the outside of the curve stretches, but the inside is compressed. The earthquakes in northern India follow exactly this behaviour, with extensional motion at shallow depths and compression deeper down (Fig. 1b). In the simplest case, the point of crossover between these two different motions — called the neutral surface — marks the middle of the beam. In India, the neutral surface turns out to be almost exactly at the middle of the crust.

The conclusion, says Jackson¹, is obvious: the strength of the Indian lithosphere must lie in the crust. And the relation between beam thickness, crustal thickness and earthquakes in other continents studied by McKenzie and Fairhead⁶ seems to point the same way. Jackson is not surprised because he thinks that the underlying mantle contains water, but that the lower crust above it is dry. In this case, given the likely temperatures in the lithosphere, laboratory experiments on rocks suggest that the mantle is weaker than the crust.

But there are some chinks in the armour of Jackson's argument. For example, the measured depth of the neutral surface in India could be misleading because the stresses induced by bending, but also the sideways push of the Himalayas and Tibet. These extra stresses will raise the level of the neutral surface in the effective elastic beam. If the neutral surface really is at the middle of

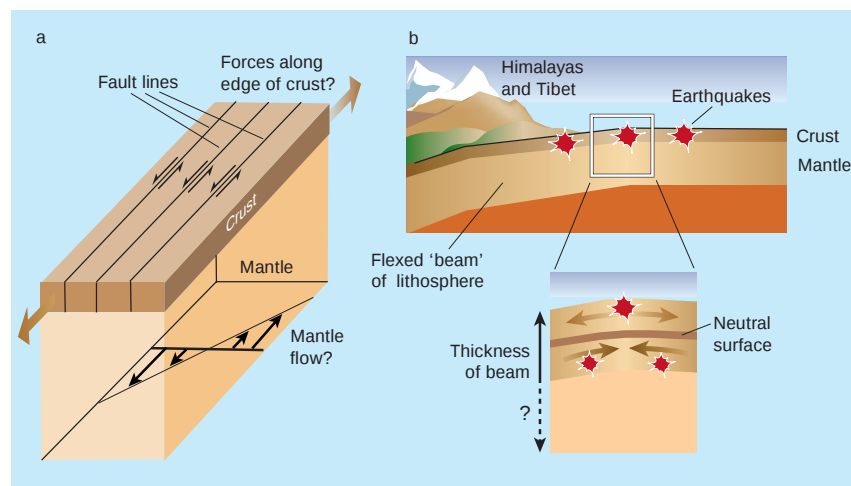


Figure 1 Continental conundrum. a, Are the forces that move fault lines on land in the Earth's crust, as one block rubs against another, or do they lie much deeper in the underlying mantle? b, Jackson¹ argues that the beam-like behaviour of northern India, flexed under the weight of the Himalayas and Tibet, may provide the answer. The top of the bent beam is stretched, whereas deeper down it is compressed: earthquakes are confined to the crust and show exactly this pattern of deformation, so at least part of the beam must lie there too.

the crust (at a depth of about 25 km), then the underlying part of the beam could be much thicker, extending into the mantle.

A more fundamental problem is that geologists have surprisingly little to go on when it comes to understanding the mantle, and there is an embarrassingly large gap between laboratory experiments and the behaviour of rocks under real geological conditions. To compound the problem, McKenzie and Fairhead's results⁶ fly in the face of scores of other studies: studies conducted in northern India and elsewhere suggest much greater values of beam thickness, up to several times the thickness of the crust, plunging geophysicists into an intense argument about the significance of this discrepancy⁷.

In any case, the outer part of the Earth is almost certainly not a single elastic beam, but is more likely to be formed from a number of distinct and separated horizontal layers. And it can be argued that where there are load-bearing regions, the occurrence of earthquakes doesn't necessarily follow — rocks can be plastic, even viscous, rather than brittle, and still offer considerable resistance to deformation. Indeed, earthquakes, in which rocks really are breaking, could be

taken as a sign of intrinsic weakness rather than just excessive strain. And a wet mantle remains as speculation: if instead it's dry, it could actually have considerable strength.

Jackson¹ has bravely stuck his neck out and triggered a vociferous but vital debate. The way forward, once geophysicists have settled their differences about measuring elastic-beam thicknesses, must lie in a renewed effort to find out what the Earth's mantle is really like. The study of rare chunks of mantle rock, brought up in volcanoes, combined with the development of better ways to measure mantle temperatures, may prove critical in this respect. Looking below the crust is not easy, but the rewards in terms of understanding our planet will be worth the effort. ■

Simon Lamb is in the Department of Earth Sciences, University of Oxford, Parks Road, Oxford OX1 3PR, UK.

e-mail: simon.lamb@earth.ox.ac.uk

1. Jackson, J. A. *GSA Today* **12** (9), 4–9 (2002).
2. England, P. C. & Molnar, P. *Science* **278**, 647–650 (1997).
3. Lamb, S. H. *J. Geophys. Res.* **99**, 4457–4483 (1994).
4. Maggi, A. *et al. Geophys. J. Int.* **143**, 629–661 (2000).
5. Maggi, A. *et al. Geology* **28**, 495–498 (2000).
6. McKenzie, D. P. & Fairhead, D. *J. Geophys. Res.* **102**, 27523–27552 (1997).
7. Watts, A. B. *Isostasy and Flexure of the Lithosphere* (Cambridge Univ. Press, 2001).

Applied physics

Terahertz power

Mark Sherwin

Although radiation at terahertz frequencies has many uses, most sources cannot generate terahertz beams with great power. Magnetic manipulation of energetic electrons inside a particle accelerator offers a solution.

This page emits a broad spectrum of electromagnetic radiation, including frequencies in the region of one terahertz (1 THz = 10^{12} Hz). You cannot see this terahertz emission because its frequency is about 300 times smaller than the limit of human vision. Neither can you feel it: the total intensity emitted at all frequencies below 1 THz is less than a millionth of a watt per square centimetre. Not only this page, but all of the objects around you emit terahertz electromagnetic waves in all directions as 'black-body radiation'.

Terahertz beams much brighter than black-body radiation are required for scientific and technological applications, ranging from the imaging of biological and other materials^{1,2} to manipulating quantum states in semiconductors³. On page 153 of this issue, Carr *et al.*⁴ report the generation of a beam of radiation that contains a broad spectrum of frequencies up to about one terahertz with an average power of 20 W. Such a beam has never previously been created; Carr *et al.* have opened the door to new investigations and applications in a wide range of disciplines.

Researchers have at their disposal an

increasing number of sources of coherent terahertz radiation — that is, terahertz radiation with a well-defined phase, such that it can be tightly focused. Lasers that can generate pulses of visible or near-infrared light (around 10^{14} – 10^{15} Hz) with a duration less than 10^{-12} s are increasingly common, and can, with small incremental costs, be used to generate terahertz radiation⁵.

One common method is as follows. An electric field of about 10^6 V cm⁻¹ is generated in a high-resistance semiconductor by applying a d.c. voltage between a pair of electrodes bonded to its surface. An ultrafast laser pulse illuminates the semiconductor between the electrodes, creating a large density of mobile charge carriers (electrons and 'holes') through an effect that is closely related to the photoelectric effect used in solar cells. These charge carriers, sensing the large electric field, accelerate at roughly 10^{17} m s⁻² — compare that to the gravitational acceleration felt by an object dropped near the Earth's surface, of 10 m s⁻². All accelerating charges emit electromagnetic radiation. These charge carriers, reaching their maximum velocity in less than 10^{-12} s,

emit a single electric-field pulse shorter than 10^{-12} s that contains a broad range of frequencies, up to a few terahertz. Typically, the average power generated by this method is less than 10^{-6} W. But as this power is in a stable, coherent beam with well-known temporal characteristics, it can be used for spectroscopy with high spectral resolution and excellent signal-to-noise ratio, and even for imaging³. The drawback in imaging, however, is that it is usually necessary to scan the beam spot over the object in question, which is much too slow for video-rate image acquisition.

Carr *et al.*⁴ also use accelerating electrons to generate their 20-W beam of broadband terahertz radiation. But rather than being trapped inside a semiconductor, these electrons are travelling in a vacuum at nearly the speed of light — inside an accelerator at the Jefferson Laboratory in Newport News, Virginia. The electrons are grouped in bunches that are so small they whiz past an observer in 0.5×10^{-12} s. As long as a bunch of electrons travels in a straight line, it does not accelerate or emit light. But a strong magnetic field can deflect the bunch: if its trajectory is bent along a circular arc of radius 1 m, the associated acceleration causes the bunch to emit a 500-fs pulse (1 fs = 10^{-15} s) of electromagnetic radiation with a peak power of roughly 10^6 W, a peak frequency of about 0.6 THz, and detectable radiation up to several terahertz. When electron bunches are generated at the maximum rate of 37 million each second, the average power in the beam reaches roughly 20 W. In fact, the authors even needed to reduce the power by a factor of 550 to bring the generated signal back within the dynamic range of their equipment.

The 20-W broadband beam complements other sources of terahertz radiation. These include the broadband sources based on ultrafast lasers, discussed above, that come in two types. The first emits pulses with peak powers that are similar to those reported by Carr *et al.*⁴, but the maximum repetition rate is only 10^3 Hz, compared with 4×10^7 Hz for the Jefferson Lab source, and the resulting average power is roughly 10^{-3} W (ref. 6). The other type of source emits terahertz pulses with a repetition rate of 10^8 Hz, but an average power of less than 10^{-6} W (refs 1, 3).

There are also sources that emit radiation at well-defined frequencies. These include the quantum-cascade laser⁷, which produces pulses with a peak power of 0.002 W at around 4 THz; microwave oscillators⁸, whose frequency can be multiplied by non-linear devices (10 s of continuous power at the microwatt level and below 2 THz); and free-electron lasers that reach 10^3 – 10^6 W of peak power⁹. Each of these sources has already found its own applications and users.

As with any new technology, it is difficult

to predict the most important applications — the inventors of the laser did not envisage barcode scanners. The authors speculate that the large peak power could be used for the study of new nonlinear phenomena in advanced materials and devices, and that the large average power could allow “full-field, real-time image capture” — in effect, terahertz movies. Another possibility is that the large average and peak powers could be used to manipulate and alter materials, chemical reactions and biological processes. Perhaps, as you bask in the weak terahertz radiation emitted by this page, you will think up the ‘killer’ application for the new terahertz source. ■

Mark Sherwin is in the Department of Physics,

University of California, Santa Barbara, California 93106, USA.

e-mail: sherwin@physics.ucsb.edu

1. Hu, B. B. & Nuss, M. C. *Optics Lett.* **20**, 1716–1719 (1995).
2. Chen, Q. & Zhang, X.-C. in *Ultrafast Lasers: Technology and Applications* (eds Fermann, M. E., Galvanauskas, A. & Sucha, G.) 521–572 (Dekker, New York, 2001).
3. Cole, B. E., Williams, J. B., King, B. T., Sherwin, M. S. & Stanley, C. *Nature* **410**, 60–63 (2001).
4. Carr, G. L. *et al.* *Nature* **420**, 153–156 (2002).
5. Grischkowsky, D., Keiding, S., van Exter, M. & Fattinger, C. *J. Opt. Soc. Am. B* **7**, 2006–2015 (1990).
6. You, D., Jones, R. R., Bucksbaum, P. H. & Dykaar, D. R. *Optics Lett.* **18**, 290–292 (1993).
7. Köhler, R. *et al.* *Nature* **417**, 156–159 (2002).
8. Martin, S. *et al.* in *2001 IEEE MTT-S Int. Microwave Symp. Digest*, Vol. 3 (ed. Sigmon, B.) 1641–1644 (IEEE, Piscataway, New Jersey, 2001).
9. Colson, W. B. *et al.* *Physics Today* **55** (1), 35–41 (2002).

Neuroscience

Single-neuron mnemonics

Barry W. Connors

How can you remember what you've just read or seen or done? The issue of short-term memory has vexed neuroscientists for more than half a century; a new study adds an unexpected piece to the puzzle.

Even the minutiae of life must be remembered, however fleetingly. Who just called me? Which journal am I reading? What were the words that preceded this question mark? And as new challenges appear, it also helps to forget the trivia of the recent past. The capacity to temporarily retain small quantities of information is loosely called short-term or working memory, and is a basic function of the brain. Yet the mechanisms underlying this ability remain a mystery. Long-lasting forms of memory

apparently require molecular or structural changes in nerve cells, but short-term memory is a dynamic, ephemeral process that has not yielded to molecular characterization. A venerable theory is that short-term memories are held by interconnected groups of neurons that fire persistently because they excite one another recursively. The work of Egorov and colleagues¹, described on page 173 of this issue, thickens the plot. These authors found that a single, isolated neuron, when stimulated briefly, could generate

sustained increases in its electrical activity that were graded in intensity and readily reversible. In other words, one such neuron could quickly remember (and forget) numerous bits of information.

Any theory of short-term memory must explain how it is initiated, why it persists, what makes it specific, and how it ends. Individual neurons had previously seemed too inflexible for the job. Most neurons, when isolated from the rest of the brain, increase their activity only during a stimulus (Fig. 1a), so it's hard to see how they could be involved in remembering that stimulus. A few neurons are bistable — brief stimuli can switch them between two persistent states (Fig. 1b) — and, like electronic flip-flops, such neurons can in principle serve as memory media². But they are exceedingly rare in mammals.

Moreover, in the brains of animals that are performing certain tasks, neurons seem to do something even more complex³. In a typical experiment, a monkey is trained to remember a short sensory cue, for example a coloured flash of light, then wait patiently during a variable delay, and finally respond to the remembered cue in a manner appropriate to the task. Groups of neurons in many areas of the brain's cerebral cortex both increase and sustain their firing rates throughout the delay period, suggesting that they are involved in remembering the stimulus. Such ‘delay neurons’ often have distinct stimulus preferences, persistently firing faster after the cue is, say, green rather than red.

How might such delay activity come about? In the 1930s, Rafael Lorente de Nó pointed out⁴ the dense interconnections between neurons throughout the brain. He suggested that a brief stimulus might excite a few neurons, which could excite other neurons, such that neural activity ricochets around, or reverberates, even after the original stimulus has disappeared (Fig. 1d). The reverberations would probably be fragile and easily disrupted by another salient stimulus. Although born in the primordial ooze of theoretical neuroscience, the notion of persistently active neuronal networks — famously elaborated by Karl Lashley and Donald Hebb⁵ in the 1940s — still holds sway. Contemporary theoreticians have modelled formal, clever variations of it that explain a range of specific behaviours^{6–9}. In most models each neuron is given mundane intrinsic properties (such as those seen in Fig. 1a), and delay activity emerges only as a consequence of the interactions among many neurons.

But the findings of Egorov *et al.*¹ suggest that this is not the only possible explanation. These authors studied a region of the rat brain called the entorhinal cortex, which is essential for working memory during certain behaviours. They first exposed slices

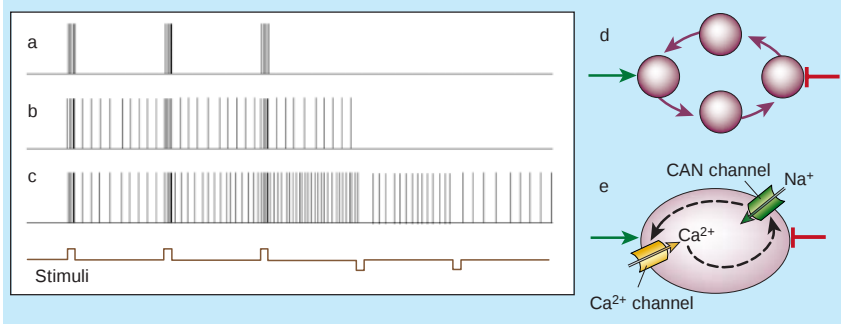


Figure 1 Memory and the single neuron. Neurons signal their activity by generating brief spikes of membrane voltage; the rate and timing of spikes carry information. **a**, Most neurons respond to excitatory stimuli (upward steps in the line below **c**) by spiking only as long as each stimulus lasts. **b**, Very rare neurons are bistable: brief excitation leads to persistent spiking, always at the same rate; brief inhibition (downward steps in the line below **c**) can turn it off. **c**, Multistable neurons persistently increase or decrease their spiking across a range of rates in response to repeated brief stimuli. **d**, In a reverberatory network model of short-term memory, an excitatory stimulus (green arrow) leads to recursive activity (purple arrows) in interconnected neurons (purple circles). Inhibitory stimuli (red bar) can halt the activity. **e**, Egorov *et al.*¹ suggest that graded persistent activity in single neurons (as in **c**) might be triggered by a pulse of internal Ca^{2+} ions that enter through voltage-gated channels; Ca^{2+} then activates CAN channels, through which an inward current (largely comprising Na^{+} ions) enters, persistently exciting the neuron. The positive feedback loop (black broken arrows) may include numerous, unknown signalling mechanisms.

of cortex to a drug that mimics acetylcholine — a natural substance that facilitates particular memory functions — and then applied a brief excitatory stimulus to a single neuron or to the slice. Single entorhinal neurons responded by increasing their firing rate for a minute or more. Surprisingly, when similar stimuli were repeated, the firing rate increased in a stepwise manner; short, inhibitory stimuli progressively decreased the rate (Fig. 1c). And the neurons could do all this even when they were effectively disconnected from other neurons. Their behaviour is reminiscent of that of the delay neurons in the brains of animals performing working-memory tasks and in network models of short-term memory.

How can one neuron do what had seemed to be the exclusive province of neuronal networks? Egorov *et al.* provide a few clues. They show that the triggering stimulus must cause Ca^{2+} ions to enter the neuron, probably through Ca^{2+} -selective membrane channels (Fig. 1e). Internal Ca^{2+} regulates countless cellular processes, but here it seems to activate Ca^{2+} -dependent nonspecific cation (CAN) channels, which provide an inward ionic current that further excites the neurons. By this scheme, Ca^{2+} is the trigger stimulus and the CAN current underlies the persistent activity.

But there are still major puzzles. What accounts for the graded nature of the persistent activity? How can it be quickly reversed? And what maintains the delicate, sustained balance of excitability? There are many possibilities. Perhaps each pulse of Ca^{2+} triggers (directly or indirectly) a change of state in just a fraction of the CAN (or some other) channels, slightly enhancing the neuron's excitability. During brief periods of inhibition, internal Ca^{2+} levels might fall, and the CAN channels would relax towards their resting state. A less conventional scenario is that some of the neuron's dendrites (long, thin membrane extensions) exhibit bistable electrical behaviour, for which Ca^{2+} channels¹⁰ and CAN channels are responsible. Brief stimuli might then activate or inactivate each dendrite successively, yielding multistable states in the neuron as a whole. Some of these ideas could be tested by imaging dendritic Ca^{2+} levels during persistent activity.

The burning question is, does the single-neuron behaviour described by Egorov *et al.* really play a role in short-term memory? That remains to be seen. Individual neurons are unlikely to go it alone, because memories are distributed across large numbers of neurons³. But perhaps intrinsically mnemonic neurons are an essential component of interconnected networks that encode memories¹¹. The neurons and networks described here all have a feature that is essential for persistent firing: excitatory feedback that maintains neural activity after a short

stimulus gets it started. In the networks, such positive feedback is provided by interconnections between neurons. In the single neuron, feedback is provided by the internal Ca^{2+} concentration, the inward current it regulates, the cell's membrane voltage, and undoubtedly other components. One of the tricky problems in most formal memory models is stability. Unless the timing, strength and duration of the positive feedback are precisely tuned, memories may oscillate, persist inappropriately, or fade too quickly⁸. The stability of working memory might be optimized when intrinsically multistable neurons are elements of interconnected networks.

Without short-term memory, cognition itself crumbles. Disorders of working memory have, for instance, been implicated in such devastating psychiatric diseases as schizophrenia. If a single-neuron mnemonic mechanism does prove relevant to behav-

iour, it will help us to understand working memory — and its dysfunctions — at the molecular scale.

Barry W. Connors is in the Department of Neuroscience, Brown University, Providence, Rhode Island 02912, USA.

e-mail: barry_connors@brown.edu

1. Egorov, A. V., Hamam, B. N., Fransén, E., Hasselmo, M. E. & Alonso, A. A. *Nature* **420**, 173–178 (2002).
2. Marder, E., Abbott, L. F., Turrigiano, G. G., Liu, Z. & Golowasch, J. *Proc. Natl Acad. Sci. USA* **93**, 13481–13486 (1996).
3. Goldman-Rakic, P. S. *Neuron* **14**, 477–485 (1995).
4. Lorente de N6, R. J. *Neurophysiol.* **1**, 207–244 (1938).
5. Orbach, J. *The Neuropsychological Theories of Lashley and Hebb* (Univ. Press America, Lanham, Maryland, 1998).
6. Amit, D. J. & Brunel, N. *Cereb. Cortex* **7**, 237–252 (1997).
7. Lisman, J. E., Fellous, J.-M. & Wang, X. J. *Nature Neurosci.* **1**, 273–275 (1998).
8. Seung, H. S., Lee, D. D., Reis, B. Y. & Tank, D. W. *J. Comput. Neurosci.* **9**, 171–185 (2000).
9. Durstewitz, D., Seamans, J. K. & Sejnowski, T. J. *Nature Neurosci.* **3**, 1184–1191 (2000).
10. Reuveni, I., Friedman, A., Amitai, Y. & Gutnick, M. J. *J. Neurosci.* **13**, 4609–4621 (1993).
11. Fransén, E., Alonso, A. A. & Hasselmo, M. E. *J. Neurosci.* **22**, 1081–1097 (2002).

Solar physics

The Sun under a microscope

John H. Thomas

Fine details of the filamentary structure of sunspots are revealed in new observations. These high-resolution measurements herald the quality of data to be expected from a new generation of solar telescopes.

Many scientific questions about the Sun can only be answered by observations made with high spatial resolution^{1,2}. At the solar surface, turbulent convection and magnetic fields interact in a complicated manner, generating fine-scale structures. This is especially evident in a sunspot, with its dark central core, or umbra,

surrounded by a region of radial filaments, known as the penumbra. High-resolution observations have revealed much of the intricate thermal and magnetic structure of the penumbra, but have also hinted at structure on scales below the resolution limit of existing telescopes³.

On page 151 of this issue, Scharmer *et al.*⁴

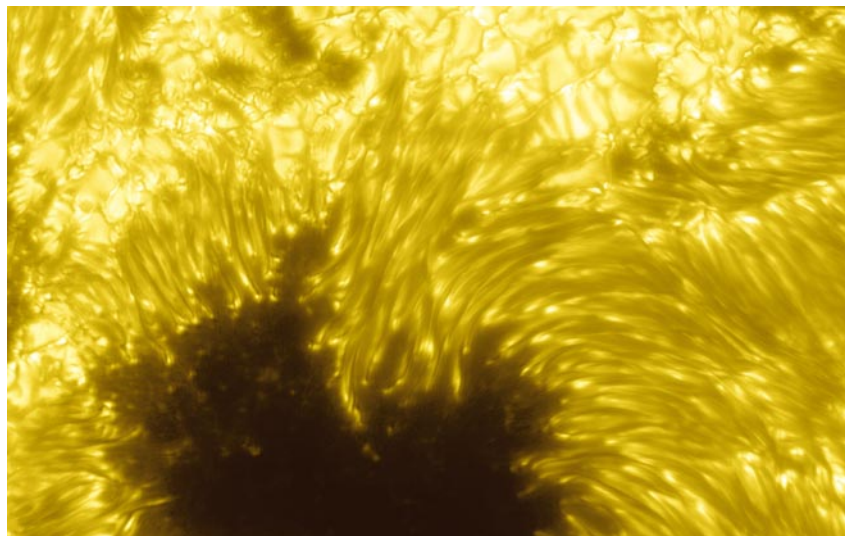


Figure 1 Darkness in the light. Around the dark core of a sunspot, a filamentary structure develops, driven by intense magnetic fields and gas flow. This image from Scharmer *et al.*⁴ shows in unprecedented detail the structure of the filaments themselves, in particular their dark central cores, for which there is as yet no explanation.

report new observations of the penumbral filaments with unprecedented spatial resolution (Fig. 1). They have detected dark central cores, less than 90 km wide, within many of the filaments (which are typically 150–180 km wide). This discovery is the first scientific result from the new Swedish 1-m Solar Telescope, which became operational in May 2002.

It is appropriate that this first result is for sunspots, which have long been a test bed for studying magnetohydrodynamics — how conducting fluids, such as plasmas or liquid metals, behave in magnetic fields — under astrophysical conditions. The filamentary penumbra has proved especially difficult to understand, with its complicated magnetic-field geometry and associated gas flows and oscillations^{5,6}. The filament cores discovered by Scharmer *et al.*⁴ are likely to be an important key to understanding the penumbra. But their results are based on narrow-band imaging, which reveals only the intensity (thermal) pattern of the dark cores. As the authors point out, a full understanding of these features will require measurements of the magnetic field and flow velocity in the cores using the more difficult techniques of spectroscopy and polarimetry.

An essential feature in these new observations is the use of an adaptive optics (AO) system, operating in real time to correct for the distorting effects of the Earth's atmosphere on sunlight. In an AO system, a reference image is used to detect the distortions, and then the sections of a multi-segment mirror are moved independently to correct for them. The system used by Scharmer *et al.* at the Swedish 1-m Solar Telescope has 19 mirror elements. In contrast to night-time observations, where a steady, point-like source (a single star) is available as a reference, solar AO systems have solved the more difficult problem of how to use an extended, low-contrast, time-varying image such as a solar granulation pattern as a reference. An AO system is essential for any large ground-based solar telescope to achieve consistent high resolution near the limit of its power (its diffraction limit).

Adaptive optics systems have breathed new life into some older solar telescopes. A 24-element AO system has been in use for three years at the Dunn Solar Telescope at the US National Solar Observatory (NSO) in Sunspot, New Mexico. Other AO systems are also in use at the German Vacuum Tower Telescope on Tenerife in the Canary Islands and at the largest existing solar telescope, the NSO's McMath–Pierce Telescope on Kitt Peak in Arizona, where sensitive measurements of solar magnetic fields are being made at near-infrared wavelengths. An 80-element AO system is under development for use both at the Dunn telescope and at the Big Bear Solar Observatory in California.

High-resolution imaging in solar astron-

omy relies not only on angular resolution but also on spectral (wavelength) and temporal resolution. Good angular resolution typically reveals only the horizontal surface structure of the Sun. To sample different heights in the solar atmosphere, measurements at different wavelength positions within a spectral absorption line are needed. In general, the smaller the solar feature, the more rapidly it changes with time. So, as angular resolution is increased, temporal resolution must also be increased to follow these changes. To achieve high angular, temporal and spectral resolution simultaneously requires a high throughput of light, and this may well dictate a larger telescope aperture than would angular resolution alone.

The need for large aperture is a driving factor in the design of new solar telescopes. A German consortium is reconfiguring a telescope on Tenerife into the GREGOR telescope with a 1.5-m aperture, with completion expected in 2005. The most ambitious project is the Advanced Technology Solar Telescope (ATST) at the NSO. The ATST will have a 4-m aperture, which will provide high angular resolution (less than 0.1 arcsecond) at wavelengths where very sensitive magnetic measurements can be made. The ATST will also observe out to relatively long wavelengths in the thermal infrared region, where magnetic fields in the solar corona can be measured. Design studies and site testing for the ATST are now being carried out, and first light is scheduled for 2010. Of course, the ideal high-resolution instrument would be a large solar telescope in space — but the largest project at present is a telescope with a relatively small aperture of 50 cm, scheduled for launch on the Japan–US–UK Solar-B satellite in 2005.

Is there a limit to how small a structure

we might expect to observe on the Sun? Theoretical arguments suggest that there are magnetic structures on scales as small as a kilometre, or less, in the Sun's photosphere (its visible surface). Computer simulations of photospheric magnetoconvection show very small structures, but the simulations have not yet achieved sufficient resolution to determine the limiting size. The horizontal mean free path — in other words, the average distance travelled without interacting — of a photon in the solar photosphere is about 50 km, and so this might be expected to be the smallest observable length scale, because of the smoothing effect of radiative energy transfer. But sophisticated radiative-transfer calculations show that fine structures as small as a few kilometres should in principle be directly observable⁷. It seems, then, that the quest for higher resolution in solar astronomy — and with it, the increase in understanding of our own star — is likely to continue well beyond the telescope projects now under way. ■

John H. Thomas is in the Departments of Physics & Astronomy and Mechanical Engineering, University of Rochester, Rochester, New York 14627, USA.
e-mail: thomas@astro.me.rochester.edu

1. Rimmele, T. R., Balasubramaniam, K. S. & Radick, R. R. (eds) *ASP Conf. Series 183, High Resolution Solar Physics: Theory, Observations, and Techniques* (Astron. Soc. Pacific, San Francisco, 1999).
2. Thomas, J. H. *Nature* **396**, 114–115 (1998).
3. Sánchez Almeida, J. & Lites, B. W. *Astrophys. J.* **532**, 1215–1229 (2000).
4. Scharmer, G. B., Gudiksen, B. V., Kiselman, D., Löfdahl, M. G. & Rouppe van der Voort, L. H. M. *Nature* **420**, 151–153 (2002).
5. Thomas, J. H. & Weiss, N. O. (eds) *Sunspots: Theory and Observations* (Kluwer, Dordrecht, 1992).
6. Schmieder, B., del Toro Iniesta, J. C. & Vázquez, M. (eds) *ASP Conf. Series 118, Advances in the Physics of Sunspots* (Astron. Soc. Pacific, San Francisco, 1997).
7. Bruls, J. H. M. J. & von der Lühse, O. *Astron. Astrophys.* **366**, 281–293 (2001).

Medicine

Lipid signals in pain control

Nicolas G. Bazan and Rod J. Flower

Cyclooxygenase enzymes produce lipid messenger molecules whose roles in health or disease depend on their context. The discovery of cyclooxygenase-3 should enhance our knowledge of such events.

Some of the most widely used medicines today are aspirin and other such nonsteroidal anti-inflammatory drugs (NSAIDs), which are well known for their pain-relieving, fever-reducing and anti-inflammatory effects. Simply put, these drugs work mainly by inhibiting the formation of prostaglandins — potent lipid messenger molecules — and they do this by blocking cellular cyclooxygenase enzymes, of which two have previously been discovered. It has long been a puzzle, however, that acetaminophen, otherwise known

as paracetamol, works well to relieve pain and reduce fever, yet is not an effective anti-inflammatory medicine (although it is often classified as an NSAID). Also, it does not markedly inhibit the two known cyclooxygenases. As they describe in *Proceedings of the National Academy of Sciences*, Daniel Simmons and colleagues¹ might have found part of the solution: they show that acetaminophen inhibits a third cyclooxygenase, which is strongly expressed in the brain and may be involved largely in mediating pain.

The membranes of our cells are a Pandora's

Cyclo-oxygenase	Gene of origin	Expression characteristics	Functions	Inhibited by
COX-1	COX-1	Constitutively expressed	Involved in organ pain, platelet activation, stomach protection	NSAIDs, including aspirin; other pain-relieving drugs
COX-3	COX-1	Constitutively expressed, particularly highly in brain	Possibly involved in sensing pain; little involvement in inflammation	Highly sensitive to acetaminophen; also inhibited by some NSAIDs
PCOX-1a	COX-1	Not known	Not known	Not known
PCOX-1b	COX-1	Not known	Not known	Not known
COX-2	COX-2	Induced by growth factors, neurotransmitters, inflammatory cytokines, oxidative stress, injury, inflammation, low oxygen levels in the brain, seizures, neurodegeneration. Constitutively expressed in brain, kidneys, and other organs	Inducible COX-2 causes inflammatory pain and fever; constitutive COX-2 is involved in synaptic plasticity	NSAIDs; COX-2-selective drugs

Figure 1 The growing family of cyclooxygenases. The figure shows the known cyclooxygenases (COXs), some of their known sites of expression and functions, and the drugs (including nonsteroidal anti-inflammatory drugs, NSAIDs) that inhibit them. Simmons and colleagues¹ have discovered that several enzymes are encoded by the COX-1 gene, including COX-3 and two small partial COX-1 proteins, PCOX-1a and PCOX-1b. COX-3 is very sensitive to acetaminophen.

box of lipid messenger molecules, many of which have powerful effects — some good, some bad — on the body. Enzymes such as phospholipase A₂ play the role of Pandora in releasing these messengers, by splitting a certain chemical bond in phospholipids, the predominant membrane components. Phospholipase A₂, for instance, releases arachidonic acid from phospholipids; arachidonic acid in turn serves as a substrate for the cyclooxygenases, which generate a short-lived intermediate, prostaglandin H₂. This lipid is the precursor to a variety of derivatives, which share family resemblances in their chemical structures, but do different things. For example, some make nerve endings hypersensitive and others lead to inflammation or fever. Some activate platelets, thereby contributing to blood clotting, and others help protect the stomach lining. Yet more participate in birth, bone remodelling, and sleep induction; some modulate neurotransmitter release; some are anticonvulsive; and others might participate in Alzheimer's disease.

Back in 1971 only one cyclooxygenase was known, and Vane² discovered that it is inhibited by NSAIDs. This finding provided a unifying explanation for the fever-reducing, pain-relieving and anti-inflammatory actions of NSAIDs, and for the fact that these drugs have qualitatively similar side-effects, such as preventing blood clotting and harming the gastrointestinal tract. The results also firmly established the prostaglandins as critical mediators of inflammatory disease.

But the discovery could not explain why different NSAIDs, at similar doses, have quantitatively different effects and side-effects. A mystery surrounded acetaminophen in particular. Hugely popular, with a history almost as venerable as that of aspirin, this drug reduces pain and fever but has little

anti-inflammatory activity, and has virtually no side-effects on the stomach and platelets (although there have been concerns about its effects on the liver). It was suggested that acetaminophen specifically affects the brain, which mediates pain and controls fever, and the general idea developed that there are various — perhaps tissue-specific — cyclooxygenases that are affected differently by different NSAIDs³.

That idea lay dormant until the early 1990s, when, amid great excitement, researchers discovered^{4,5} a second cyclooxygenase, COX-2. The original enzyme, now dubbed COX-1, is expressed constitutively, including in platelets and the stomach (Fig. 1). By contrast, COX-2 is generally expressed under specific circumstances, largely as a result of inflammation: it is induced by cytokine proteins associated with inflammation, and by injury, seizures, and so on⁶. However, in the brain, kidneys and some other tissues, COX-2 is expressed constitutively, illustrating the multifunctional nature of these enzymes. For example, the brain COX-2 gene is expressed as a result of synaptic activation in a totally physiological event⁷.

It also soon became clear that different NSAIDs have different effects on these two cyclooxygenases, and that this could explain the drugs' quantitatively different effects. For instance, drugs that inhibit COX-2 strongly but COX-1 weakly show potent anti-inflammatory activity, with fewer gastrointestinal side-effects⁸. Some of these drugs — etodolac, meloxicam and nimesulide — already existed, but were now found to inhibit COX-2 preferentially. Others were developed as selective COX-2 inhibitors, including rofecoxib and celecoxib.

Yet the discovery of COX-2 did not solve the acetaminophen mystery. At the concen-

trations used therapeutically, this drug inhibits both COX-1 and COX-2 only weakly. This suggested the existence of yet more cyclooxygenases, and a few years ago Simmons and colleagues⁹ discovered, in cell lines, some forms of COX-2 with slightly different properties. Now the Simmons group¹ has discovered an entirely new cyclooxygenase, COX-3, which they believe to be the acetaminophen-sensitive species. They have also found two smaller forms of COX-1.

All of these are derived from the COX-1 gene by 'alternative splicing' of the COX-1 messenger RNA. So, to make the mature mRNA encoding COX-1 protein, cells remove a particular stretch of sequence, intron 1. To make the mature mRNA for COX-3, intron 1 is included. The same is true of the mRNA encoding one of the small COX-1 derivatives, although here some further sequences are removed. COX-3 therefore has the same sequence (except for intron 1) as COX-1, and both are bound to the cell membrane. The retention of intron 1 might change the way the enzyme molecule is folded, or the conformation of its active site.

Simmons and colleagues¹ also found that, in humans, COX-3 is most abundant in the heart and the cerebral cortex, where it is present at about 5% of COX-1's concentration. *In vitro*, COX-3 is more sensitive than COX-1 or COX-2 to inhibition by acetaminophen, and is much more sensitive to diclofenac, indomethacin, ibuprofen and aspirin — the NSAIDs tested so far. So it seems plausible that the pain-relieving effects of these drugs are mediated largely through the nervous-system-located COX-3. It's still not clear, however, how acetaminophen reduces fever. When the COX-2 gene is knocked out in mice, the drug no longer has fever-reducing effects, which might suggest that it works through the COX-2 protein to reduce fever. Yet concentrations of acetaminophen that reduce fever *in vivo* do not appear to inhibit COX-2. Perhaps, like the COX-1 gene, the COX-2 gene can also produce different enzymes, one of which is affected by acetaminophen. Simmons and colleagues¹ also suggest that there may be other COX enzymes.

Why are there several different cyclooxygenases? They all seem to have broadly similar structures, and catalyse the same reaction. And how can they have such different effects? COX-2, for instance, is expressed normally in several tissues, whereas elsewhere its production is increased rapidly in pathological conditions⁷. In the brain, prostaglandin E₂ produced by COX-2 specifically regulates the synaptic plasticity that is relevant to memory¹⁰. And phospholipase A₂ and arachidonic acid modulate excitatory neurotransmission and neuronal survival¹¹, through mechanisms that may involve a cyclooxygenase. This diversity could stem

from the fact that cyclooxygenases have substrates other than arachidonic acid, and act in different cellular locations and signalling pathways. For example, COX-2 converts anandamide and 2-acylglycerol — lipids that bind to cannabinoid receptors on specific nerve cells — into prostaglandin ethanolamines^{12,13}; this may be involved in pain mechanisms. COX-2 also selectively oxygenates *N*-arachidonoyl glycine, resulting in amino-eicosanoids¹⁴, whose exact function is unclear.

Simmons and colleagues' discovery¹ of a third cyclooxygenase will no doubt herald yet more twists and turns to this story. There is still much to learn about these enzymes, but the effort will be well worthwhile given their many roles in health and disease. ■

Nicolas G. Bazan is at the LSU Neuroscience Center and Department of Ophthalmology, Louisiana State University Health Sciences Center, 2020 Gravier Street, Suite D, New Orleans, Louisiana 70112, USA.

e-mail: nbazan@lsuhsc.edu

Rod J. Flower is at the William Harvey Research

Institute, Charterhouse Square, London

EC1M 6BQ, UK.

e-mail: r.j.flower@qmul.ac.uk

1. Chandrasekharan, N. *et al. Proc. Natl Acad. Sci. USA* **99**, 13926–13931 (2002).
2. Vane, J. R. *Nature New Biol.* **231**, 232–235 (1971).
3. Flower, R. J. & Vane, J. R. *Nature* **240**, 410–411 (1972).
4. Xie, W. L., Chipman, J. G., Robertson, D. L., Erikson, R. L. & Simmons, D. L. *Proc. Natl Acad. Sci. USA* **88**, 2692–2696 (1991).
5. Fletcher, B. S., Kujubu, D. A., Perrin, D. M. & Herschman, H. R. *J. Biol. Chem.* **267**, 4338–4344 (1992).
6. Marcheselli, V. L. & Bazan, N. G. *J. Biol. Chem.* **271**, 24794–24799 (1996).
7. Bazan, N. G. *Nature Med.* **7**, 414–415 (2001).
8. Warner, T. D. *et al. Proc. Natl Acad. Sci. USA* **96**, 7563–7568 (1999).
9. Simmons, D. L., Botting, R. M., Robertson, P. M., Madsen, M. L. & Vane, J. R. *Proc. Natl Acad. Sci. USA* **96**, 3275–3280 (1999).
10. Chen, C., Magee, J. & Bazan, N. G. *J. Neurophysiol.* **87**, 2851–2857 (2002).
11. Kolko, M., DeCoster, M. A., Rodriguez de Turco, E. B. & Bazan, N. G. *J. Biol. Chem.* **271**, 32722–32728 (1996).
12. Dinh, T. P. *et al. Proc. Natl Acad. Sci. USA* **99**, 10819–10824 (2002).
13. Walker, J. M., Huang, S. M., Strangman, N. M., Tsou, K. & Sañudo-Peña, M. C. *Proc. Natl Acad. Sci. USA* **96**, 12198–12203 (1999).
14. Prusakiewicz, J. J., Kingsley, P. J., Kozak, K. R. & Marnett, L. J. *Biochem. Biophys. Res. Commun.* **296**, 612–617 (2002).

Developmental biology

First come, first served

Rolf Zeller and Jacqueline Deschamps

It is more than a decade since the discovery that vertebrate *Hox* genes are arranged and expressed in the same order as the body parts they help to produce. New work looks at how this is achieved in fingers and toes.

Arguably one of the most interesting aspects of developmental biology concerns the *Hox* genes, which are needed for the development of all bilaterally symmetrical animals. These genes are curious, as they are grouped into clusters (of which there are four in higher vertebrates) on chromosomes, and are lined up in these clusters in the same order as they are temporally and spatially expressed during development^{1,2}. In addition to this 'collinear' order of expression, *Hox* genes display 'quantitative collinearity' in certain tissues^{3,4} — the genes at one end of the cluster (the so-called 5' end) are expressed at the highest levels, whereas their 3'-located neighbours are expressed at progressively lower levels. In other words, the position of a gene in a cluster corresponds not only to its place and time of expression, but also to its expression level. On page 145 of this issue, Kmita and colleagues⁵ describe their large-scale, highly complex genetic investigation of the mechanisms behind quantitative collinearity in developing mouse limbs. They come to several interesting conclusions.

The *Hox* gene clusters encode proteins with a common feature, a DNA-binding region known as the homeodomain, that is

needed for the proteins to control the transcription of genes that regulate specific embryonic-patterning processes. In vertebrate embryos, all four *Hox* gene clusters are expressed in a temporally and spatially collinear manner along the head-to-tail (anterior–posterior) axis. So, in general, the gene at the 3' end of each cluster is expressed first and most anteriorly to dictate the identity of anterior tissues. The gene at the 5' end is expressed finally and posteriorly, to pattern the posterior structures.

In addition to their expression and function along this main body axis, it is thought that, during vertebrate evolution, *Hox* clusters acquired new, specialized functions. For instance, altering the expression of the 5' *HoxD* genes may have been crucial to the 'invention' of digits⁶. So, in today's limbed vertebrates, genes from the middle of the *HoxD* cluster are expressed in 'proximal' embryonic limb tissues, which will become the upper arms and upper legs; more 5' genes are expressed later, in more 'distal' tissues. Accordingly, the genes at the extreme 5' end of the cluster, *Hoxd10* to *Hoxd13*, are expressed in the most distal tissues — the future fingers and toes. Intriguingly, however, here these genes are expressed in

the reverse order to normal, and in a quantitatively collinear manner (Fig. 1a). The 5'-most gene, *Hoxd13*, is expressed most anteriorly and at the highest level; *Hoxd12*, *Hoxd11* and *Hoxd10* are expressed at progressively lower levels and in more posteriorly restricted domains.

The collinear expression of *HoxD* genes along the main embryonic axis and paired appendages is generally controlled by regulatory genetic elements (enhancers) within the *HoxD* cluster⁷. But in the distal limb bud, expression of the 5' *HoxD* genes is controlled by a digit enhancer, located at least 100,000 bases 5' to the *HoxD* cluster⁷. This element regulates the differential expression of *Hoxd10* to *Hoxd13*, and is therefore critical for digit patterning. Kmita *et al.*⁵ now reveal how this remote enhancer achieves quantitative collinearity of expression of these genes.

Past studies of the 5' *HoxD* genes involved introducing foreign gene-regulatory regions (promoters), which responded to the digit enhancer and thereby interfered with the normal regulation of these genes. This was a puzzling observation, as the remote digit enhancer normally finely discriminates between the four 5' *HoxD* promoters. So Kmita *et al.* used ingenious molecular-genetic tools to manipulate the 5' part of the *HoxD* cluster in mice, and generated a series of deletions and duplications (including the normal promoters). They then compared these with mutations that simply inactivate the gene products, without deleting the genes and their promoters.

One striking observation is that inactivation of the *Hoxd13* protein disrupts digit morphogenesis, whereas deleting the entire gene restores digit patterning almost completely. Molecular analysis showed that inactivating the protein does not alter the expression of the other 5' *HoxD* genes, whereas deleting the gene results in *Hoxd12* — now the 5'-most gene of the cluster — being expressed in a *Hoxd13*-like pattern (Fig. 1b). Moreover, duplication of the genomic region encompassing *Hoxd13* to *Hoxd11* showed that only the 5'-most of the two *Hoxd13* genes was expressed at high levels. It had already been found, by swapping 'sister' *Hox* genes (such as *Hoxa3* and *Hoxd3*) between clusters⁸, that these genes functionally replace one another when placed in the same genomic environment. Kmita *et al.* now show that the function of a given *Hox* gene can be altered by changing its position in the cluster relative to a specific control region.

The authors also found that deleting *Hoxd13* and a conserved regulatory region at the 5' border of the *HoxD* cluster (region XII) totally abolishes quantitative collinearity: *Hoxd10*, *Hoxd11* and *Hoxd12* become expressed at equally high levels (Fig. 1c). But

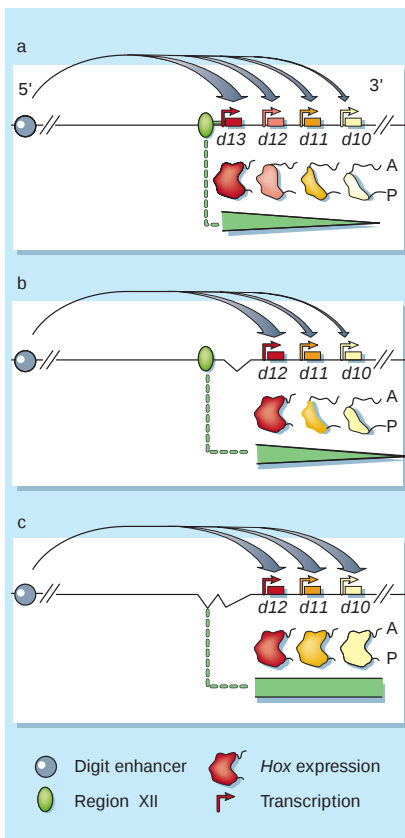


Figure 1 Regulating gene clusters in future fingers and toes. Kmita *et al.*⁵ have studied the expression of the *Hoxd13* to *Hoxd10* (*d13* to *d10*) genes in developing mouse limbs. **a**, The normal expression pattern, which results in normal digit development. These *Hox* genes were already known to be expressed in a quantitatively collinear way (green triangle), with *Hoxd13* expressed at the highest levels and most anteriorly in the future fingers and toes, and *Hoxd10* expressed least and posteriorly. A region that enhances *HoxD* transcription, the digit enhancer, is located at least 100,000 bases 5' to the gene cluster. The thickness of the black arrows shows the strength of enhancement. Kmita *et al.* have also identified another control element, region XII, that interacts with the digit enhancer. **b**, The authors show that deletion of *Hoxd13* strongly enhances *Hoxd12* expression, resulting in a *Hoxd13*-like pattern. *Hoxd12* takes over the function of *Hoxd13*. **c**, Deletion of *Hoxd13* together with region XII abrogates quantitative collinearity. A, anterior; P, posterior.

deletion of region XII alone has no effect, which reveals redundancy among the nearby regulatory elements.

During limb development, then, quantitative collinearity is established by interactions of the remote digit enhancer with regulatory elements located at the 5' border of the *HoxD* cluster. These interactions seem to have the strongest effect on the 5'-most

gene—normally *Hoxd13*—and progressively weaker effects on more 3' genes (*Hoxd12* to *Hoxd10*). It is unclear, however, whether it is strictly essential for digit development that *Hoxd13* is expressed at the highest levels and most anteriorly, and *Hoxd12* to *Hoxd10* at progressively lower levels; or whether it is simply that a leaky 'first come, first served' mechanism ensures high expression of *Hoxd13* in future digits.

The results of Kmita *et al.*⁵ are somewhat reminiscent of how the locus control region (LCR) regulates the cluster of β -globin genes⁹, which encode blood proteins that are involved in oxygen transport and storage during development and in adults. In both cases, the order of genes within the cluster is essential for their correct regulation by a remote enhancer, although the LCR activity is more sensitive to control elements that are specific to individual genes than is the digit enhancer.

It will be interesting to see whether the mechanism uncovered by Kmita *et al.* is unique to the 5' *HoxD* genes, or whether it is also used by other gene clusters to achieve new types of tissue-specific regulation. Another question is how frequently remote enhancers have been recruited to alter the expression of existing genes during the evolution of new body structures in vertebrates. One example concerns the *Sonic hedgehog* gene, which is essential for digit patterning — its limb-bud-specific expression is controlled by regulatory elements located a staggering 800 million bases 5' to the protein-coding region¹⁰. The next task will be to identify the precise molecular mechanisms that underpin long-distance concerted regulatory interactions, and the factors involved. This will be a challenge, but promises much excitement for the future.

Rolf Zeller is in the Department of Developmental Biology, Utrecht University, Padualaan 8, NL-3584CH Utrecht, The Netherlands.

e-mail: r.zeller@bio.uu.nl

Jacqueline Deschamps is in the Hubrecht Laboratory, Uppsalalaan 8, NL-3584CT Utrecht, The Netherlands.

e-mail: jacqueli@niob.knaw.nl

1. Duboule, D. & Dollé, P. *EMBO J.* **8**, 1497–1505 (1989).
2. Graham, A., Papalopulu, N. & Krumlauf, R. *Cell* **57**, 367–378 (1989).
3. Dollé, P., Izpisua-Belmonte, J. C., Falkenstein, H., Renucci, A. & Duboule, D. *Nature* **342**, 767–772 (1989).
4. Dollé, P., Izpisua-Belmonte, J. C., Brown, J. M., Tickle, C. & Duboule, D. *Genes Dev.* **5**, 1767–1777 (1991).
5. Kmita, M., Fraudeau, N., Hérault, Y. & Duboule, D. *Nature* **420**, 145–150 (2002).
6. Sordino, P., van der Hoeven, F. & Duboule, D. *Nature* **375**, 678–681 (1995).
7. Spitz, F. *et al. Genes Dev.* **15**, 2209–2214 (2001).
8. Greer, J. M., Puetz, J., Thomas, K. R. & Capecchi, M. R. *Nature* **403**, 661–665 (2000).
9. Dillon, N., Trimborn, T., Strouboulis, J., Fraser, P. & Grosfeld, F. *Mol. Cell* **1**, 131–139 (1997).
10. Lettice, L. A. *et al. Proc. Natl Acad. Sci. USA* **99**, 7548–7553 (2002).



100 YEARS AGO

The utilisation of the internal heat of the earth has often been suggested as an engineering problem of the future. The Rev. E. Rattenbury Hodges directs our attention to an issue of the *Boston News Bureau* in which a scheme is seriously proposed by the official geologist for Pennsylvania of the U.S. Geological Survey and also by Prof. Hallock, of Columbia University, New York, for drawing on the earth's internal heat by means of deep borings. The idea is to admit cold water into a deep boring and utilise the hot water and high-pressure steam produced. Mr. Hodges points out that he made similar suggestions in the *Popular Science News* for January, 1894, in an article on "Our Heat Resources of the Future". He remarked, however, at the time, "The great objection to this drawing on the earth's ancient store of thermal energy would be that her cooling and consequent shrinking would be accelerated; in other words, earthquakes would necessarily become more frequent, and possibly more violent and destructive in their effects."

From *Nature* 13 November 1902.

50 YEARS AGO

Temperature of the Interior of the Earth Bullen has derived both the density and the compressibility of the earth as functions of depth, and has found that the bulk modulus k , which is the reciprocal of the compressibility, is a linear function of the pressure p : $k = k_0 + ap$. Ramsey has re-examined this relation from the point of view of the theory of solids, and his conclusions agree with those of Bullen, namely, that the linear relation is valid both in the core and in the mantle below 1,000 km. Ramsey, however, found that the constant k_0 had different values in the core and mantle, although the derivative $dk/dp = a$ was sensibly the same... It is now possible to make an estimate of the adiabatic temperature gradient throughout the earth... Taking T at 1,000 km. to be $3,600^\circ \text{K}$, values of T can be estimated at all greater depths. It is found that the temperature at the boundary of the core and mantle is $4,350^\circ \text{K}$. and at the centre of the earth is a little greater than $4,800^\circ \text{K}$. The increase throughout the core is thus only 500° . The results should be of considerable interest to Bullard's theory of the transfer of heat from the core.

From *Nature* 15 November 1952.

Obituary

Peter W. Hochachka (1937–2002)

Peter Hochachka, a pioneer in studies of biochemical adaptation to the environment, died from cancer at his home in Vancouver on 16 September 2002. Traditionally, biochemistry and molecular biology focused on the similarities between living things. This 'unity concept' served biology well but provided a somewhat restricted view of the living world. Biologists were well aware of the wide range of temperatures compatible with life, that certain animals can live without oxygen or at extreme pressures in the deep sea. Yet the idea of biochemical adaptation to the environment was largely overlooked or even dismissed.

During his graduate-student and postdoctoral years (1959–66) at Dalhousie and Duke universities, Hochachka was subjected to the influences of enzymology and metabolic regulation as well as comparative and environmental physiology. The field of 'adaptational biochemistry' was born of the marriage between these disciplines.

In 1966, he moved to the University of British Columbia, where he spent his entire career. There, he embarked on breathtaking and original research that resulted in the development of entirely new lines of inquiry. The world was both his laboratory and his lecture hall — species, lifestyles and habitats were his variables. He led, or participated in, at least nine research expeditions on the RV *Alpha Helix*, to regions as diverse as the Amazon and the Arctic. He also took part in six expeditions to the Antarctic, four to the high Andes and one to the Himalayas. He was the biochemist in what was perhaps the first collaboration (along with specialists in engineering, anaesthesiology, paediatrics and surgery) to study metabolism in Weddell seals, free-diving from a hole in the Antarctic ice. He assembled research teams to study Andean Quechuas and Himalayan Sherpas, the first to be conducted in modern laboratories, and identified the adaptations that allow Quechuas and Sherpas to function normally at high altitude.

Hochachka brought out the best in his students and postdoctoral researchers, providing intellectual stimulation and support, as well as the freedom to be creative and to develop as independent scientists. The synergy between studies in his laboratory and those of his collaborators produced an amazing diversity of research. In the 1960s, work on enzyme adaptations led to discovery of the

evolutionary conservation of enzyme–substrate affinity and catalysed an explosion of research around the world. Investigations of enzyme adaptation to pressure uncovered the pressure-insensitivity of enzymes from deep-sea animals. Research into the tolerance of animals to anoxia began in the 1970s and resulted in the discovery of the enzyme alanine dehydrogenase and a new metabolic pathway in marine molluscs. Subsequent studies revealed that goldfish produce ethanol as an anaerobic end-product and, ultimately, led to the realization that 'good' animal anaerobes respond to anoxia by depressing their metabolic rates.

Research on bioenergetics and exercise

Innovator in studies of biochemical adaptation

included species such as rainbow trout, thoroughbred racehorses, tunas, marlin, shrews, seals, whales and hummingbirds. No hummingbird or elephant seal would have ever seen the inside of an NMR magnet had it not been for Hochachka's daring and creativity. The first of three books with George N. Somero, *Strategies of Biochemical Adaptation* (Saunders, 1973), provided the synthesis of this research and laid the groundwork for the new field of adaptational biochemistry.

Hochachka was a profound thinker with an enormous intellectual field of play. With nearly 400 publications, including seven books, his influence went far beyond adaptational biochemistry and extended into clinical medicine, exercise science, ecology and evolutionary biology. His contributions have been recognized with the highest awards that Canada can give to its scientists. In 2000, he was appointed an Officer of the Order of Canada. Later this year, he is to receive, posthumously, the Commemorative Medal marking the golden jubilee of Queen Elizabeth II. His work has helped put Canada on the global scientific map.

Hochachka will be missed for his ebullience, enthusiasm and boundless energy. To him, scientific conferences were big parties, and to socialize was to talk

about science. He wrote children's books. He made films of his research expeditions that ranged in content from action-packed sequences of tunas to the romance of monkeys holding hands in a tree while watching the African sunset. Like Albert Szent-Gyorgyi, he believed that doing science should be fun.

When he became ill with prostate cancer, this too became a matter of science. His reading of the literature led to insights into the hypoxia connection in prostate cancer-cell metabolism, which he published with his surgeons as co-authors. When a subsequent lymphoma was found to be incurable, he wrote a moving letter of thanks and farewell. In it, he said that, having consulted the powers that be, he had accepted the "assignment of examining parallel universes and the possibilities of entanglement". The curiosity that made him a great scientist and human being sustained him until his death.

Raul K. Suarez and David R. Jones
Raul K. Suarez is in the Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, California 93106-9610, USA.

e-mail: suarez@lifesci.ucsb.edu

David R. Jones is in the Department of Zoology, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada.

e-mail: jones@zoology.ubc.ca



NSERC

Halteres used in ancient Olympic long jump

These athletes worked out for themselves the optimal size of hand-held weights.

Halteres¹ (αλτήρες) are hand-held weights that were first used in the standing long jump in the eighteenth ancient Olympiad in 708 BC, and may have been introduced either to make the challenge more difficult or to extend the jumping distance². Here we use computer and experimental simulations to determine the optimal mass of halteres that would be needed to maximally extend a standing long jump, and find that this corresponds closely to the size range of actual archaeological specimens. These halteres were made of stone or lead and weighed 2–9 kg, which we calculate would increase a 3-metre jump by at least 17 cm, indicating that their purpose was to boost the performance of pentathletes. Halteres may therefore be the earliest passive tool that was devised to enhance human-powered locomotion.

Halteres were swung back and forth by the jumper before take-off, thrust forwards during the first part of the flight, and finally swung backwards just before landing, as depicted in a variety of vase paintings (Fig. 1, top; see also supplementary information)³. If we assume that a loaded body can take off at the same speed and angle as an unloaded one, the loaded jump should be longer. This is because, for the same parabolic trajectory, the body's centre of mass will be more anterior (+0.07 m) and higher (+0.08 m) at take-off, and more posterior (−0.03 m) with respect to foot contact on landing, as a result of the position of the upper limbs (for an additional mass of 3 + 3 kg).

This horizontal translation will result in a gain of 0.10 m; moreover, the parabola will be extended downwards (assuming a take-off angle of about 50°; calculated from ref. 2) to gain a further 0.07 m, giving an overall increase in jump length of 0.17 m. Throwing the halteres backwards just before landing, as suggested by some historical sources⁴, should extend this distance still further (the body plus halteres must still follow a parabolic path, so if the extra mass is thrown backwards, the rest of the body will land forwards).

However, maintenance of the same take-off speed, upon which this rationale depends, is counterintuitive for a loaded body. In a vertical jump (with swinging upper limbs), considered here as analogous to a standing long jump, the take-off speed is determined by the excess force that the muscles can exert vertically over the body's weight — so when the body is heavier (and the muscles remain the same), less 'net' force should be available to generate mechanical power. But muscles are non-



Figure 1 Use of halteres in the standing long jump. Top, vase paintings depicting the use of halteres during the middle and final parts of the flight phase of the long jump³. Bottom, archaeological halteres specimens¹ of different materials (lead, stone), shapes and masses (1.1–4.5 kg).

linear actuators, capable of producing greater force and power at lower contraction speeds⁵, and the slower forward swing of loaded upper limbs could increase the ground reaction force and affect the take-off speed and flight distance, which reflect the mechanical power produced.

To test this idea, we used a software model of the jumper (Working Model 2D, Knowledge Revolution), created from four segments assembled with realistic mass/inertia parameters and muscle-like (torque/angle and torque/angular-speed relations) actuators for both the knee⁶ and shoulder⁷ joint extensors, to simulate vertical jumps loaded with different weights in the range 0–20 kg. We found that, compared with swinging unloaded limbs, the take-off speed was 2% greater for a pair of halteres with a total mass of about 6 kg, and that the overall performance began to decline only when the extra mass exceeded 10–12 kg.

To test these predictions, we asked four subjects to perform maximal vertical jumps, during which they swung their upper limbs forwards and upwards while loaded with one pair of halteres randomly donated out of 16 that ranged from zero (unloaded) to 17 kg of total extra mass. A force platform and a motion-capture system recorded the vertical ground-reaction force and the coordinates of six body segments, from which the body's centre of mass was determined⁸.

Figure 2 shows the effects of different hand-held loads on take-off speed and wrist height. Also, the measured power peak confirms that performance is improved (by 5–7%) with a pair of halteres in the mass range 2–9 kg (5–6 kg optimal). This benefit

seems to disappear for loads over 10–12 kg.

Our results indicate that greater distance (at least 0.17 m in a 3-m jump) can be achieved during a loaded long jump, owing both to the translation of the centre of mass that results from the halteres' position, and to better exploitation of the extra muscle mass in the upper limbs, which helps to generate a greater ground-reaction force (this

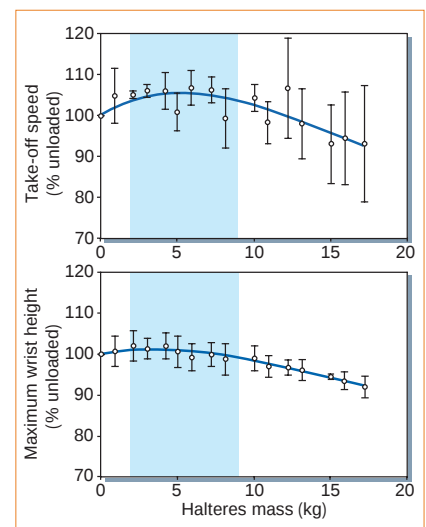


Figure 2 Effects of the mass of hand-held loads on a jumper's take-off speed and maximal wrist height during the flight. Average values ($\pm$ s.d.) of the best results (of three trials for each halteres-pair mass) are shown as a percentage of the value for an unloaded jump. Curves are third-degree polynomial regressions with imposed intercept at (0, 100%). Although the maximum wrist height shows only a slight change, this reflects a larger rise in the body's overall centre of mass. Shaded band indicates the size range of archaeological halteres specimens¹.

last argument was intuitively advanced by Aristotle⁹). The discrepancy between the predicted and actual gains in performance can be ascribed to the presence *in vivo* of elastic structures (tendons and ligaments): these can store more elastic energy because of the greater ground-reaction force, which is used successively to boost the power amplification generated by the muscle–tendon complex.

The mass range that enables all of these effects to be optimized (about 2–9 kg) corresponds closely to the actual size range of archaeological halteres specimens (Fig. 1, bottom)^{1,10}, suggesting that athletes in ancient times had worked this out for themselves.

Alberto E. Minetti, Luca P. Ardigo

Centre for Biophysical and Clinical Research into Human Movement, Department of Exercise and

Sport Science, Manchester Metropolitan University, Alsager ST7 2HL, UK
e-mail: a.e.minetti@mmu.ac.uk

1. British Museum (GR 1837.6-9.83, GR 1867.5-6.48), London, UK; National Museum, Athens, Greece; Archaeological Museum, Olympia, Greece.
2. Ebert, J. in *Abhandlungen der Sächsischen Akad. Wissensch. Leipzig VIII* 66, (Akademie, Berlin, 1963).
3. Beazley collection at www.perseus.tufts.edu
4. Wassmannsdorff, K. *Monatsschrift für das Turnwesen* 4 (1885).
5. Hill, A. V. *Proc. R. Soc. Lond. B* 127, 136–195 (1938).
6. Westing, S. H., Seger, J. Y., Karlson, E. & Ekblom, B. *Eur. J. Appl. Physiol.* 58, 100–104 (1988).
7. Perrin, D. H. *Isokinetic Exercise and Assessment* (Human Kinetics, Champaign, Illinois, 1993).
8. Minetti, A. E., Ardigo, L. P. & Saibene, F. *J. Physiol.* 472, 725–735 (1993).
9. Aristotle *Problemata* 3, 705 a, 12–19.
10. Jüthner, J. *Antike Turngeräte* (Vienna, 1896).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

COMMUNICATIONS ARISING

Evolutionary biology

Significance of primate sexual swellings

Exaggerated sexual signals are likely to be shaped by sexual selection, but few studies have examined signal evolution in females. Domb and Pagel¹ have presented support for the hypothesis that individual differences in exaggerated sexual swellings in female primates are reliable indicators of differences in female quality². However, our re-analysis of their data casts doubt on their conclusions.

In support of the sexual-selection hypothesis that swellings reliably signal enduring individual differences in female quality, Domb and Pagel¹ report that the vertical length of a female baboon's sexual swelling at maximum turgescence is correlated with female fitness. However, they did not demonstrate that variation in fertility between females is reflected in swelling size independently of other variables that might be used by males to predict reproductive output among females.

Female body height explains a significant portion of the variation in swelling size in Domb and Pagel's data³. In particular, taller females had longer swellings ($r^2=0.27$, $P=0.02$, $n=20$). In addition, the authors

pooled data from five baboon groups³. Their study groups show significant differences in swelling length ($F_{4,17}=2.96$, $P=0.05$), in age at first conception ($F_{4,24}=17.82$, $P<0.001$) and in number of offspring surviving per year ($F_{4,24}=3.93$, $P=0.014$). The groups also differed substantially in food availability³, which markedly affects female fertility⁴, and in adult sex ratio (1.3–2.2 females per male³), which might influence the male competitive regime⁵. Females in groups with higher food quality have 'longer' swellings and mature earlier. In these same groups, adult sex ratio is biased towards males, so males experience more competition per female, particularly when only one female is sexually active.

By pooling across groups, Domb and Pagel make the implicit assumption that a male can choose between females in different groups in the same manner that he chooses between females of the group in which he lives. This assumption is not supported by available data on dispersal^{6–8}. The authors controlled for age and rank in their analysis, but when female height and group membership are also controlled using their data, swelling length is no longer a significant predictor of female quality (Table 1).

The authors claim that males use the size of the sexual swelling to determine mating effort, but they did not control for adult sex

ratio — males may expend more effort simply because they have more competitors. They also analysed only the periods when a single female was receptive, excluding periods when more than one female had a swelling³. However, periods of oestrous synchrony are the best time to test whether males prefer to mate with females with larger swellings.

What do the sexual swellings of female primates indicate? We present three hypotheses that are not mutually exclusive and include that tested by Domb and Pagel. In the first, swellings are probabilistic signals of the timing of ovulation⁹; although they cannot pinpoint it¹⁰, the highest probability of ovulation is at peak swelling, making swelling size a within-cycle indicator of the probability of conception⁹. In the second, swelling size signals cycle-to-cycle variability in the probability that a female will conceive and so is a within-female indicator of the probability of conception across cycles. In the third, swelling size signals enduring differences between females in their individual ability to conceive and raise offspring, as tested by Domb and Pagel¹, and is an indicator of consistent differences in quality between females².

The within-cycle hypothesis is supported by comparative data¹. To support the between-females hypothesis and distinguish it from the within-female hypothesis, three lines of evidence are needed: differences between females in swelling size are enduring characteristics of individuals; variation between females in swelling size explains variation in fertility independently of other factors; and males prefer to mate with females with larger swellings.

However, none of this evidence has been supplied by Domb and Pagel. They have few repeated measures of female swelling size^{1,3}, precluding an analysis of whether between-female differences are consistent. Even in their small sample of repeated measures, the coefficient of variation of swelling length ranged up to 11%, which is as large as between females (11.4%; ref. 3).

We suggest that the correlation found by Domb and Pagel between swelling length and female fitness might not be due to males using individual differences in sexual swellings as an indicator of female quality, but that instead it could result from confounding variables that are correlated with swelling size. Additional information is required before alternative explanations can be ruled out. We contend that there remains no convincing empirical support for the reliable-indicator hypothesis to account for this striking signal in some primates.

Dietmar Zinner*, Susan C. Alberts†, Charles L. Nunn‡, Jeanne Altmann§

*Abteilung Verhaltensforschung & Ökologie, Deutsches Primatenzentrum, 37077 Göttingen, Germany

Table 1 Correlations between female fitness and characters

Female character	r_{length}	r_{age}	r_{rank}	r_{height}	F_{group}
Age at first conception	−0.53	0.04	0.46	0.21	6.40*
No. of offspring per year	0.39	0.26	0.24	0.11	1.27
No. of surviving offspring per year	0.46	0.48	−0.59	0.05	0.45
Proportion of offspring surviving	0.34	0.45	−0.68*	0.03	0.53
No. of offspring per year surviving ≥ 6 months†	0.25	0.29	−0.59	−0.27	2.29
Proportion of offspring per year surviving ≥ 6 months†	−0.09	0.07	−0.64*	−0.32	3.19

Correlations are shown between female fitness measurements and female characters in analyses that control for age, rank, female body size and group membership (General Regression Model, Statistica 6.0, Statsoft 2001).

Subscripted r values: length, length of swelling (swelling size); age, female age; rank, rank of family; height, height of female (body size);

group, female's group.

* $P<0.05$, two-tailed. †Analysed in ref. 3 but not included in ref. 1.

†Department of Biology, Duke University, Durham, North Carolina 27708, USA
e-mail: alberts@duke.edu

‡Section of Evolution and Ecology, University of California, Davis, California 95616, USA

§Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA

1. Domb, L. G. & Pagel, M. *Nature* **410**, 204–206 (2001).
2. Pagel, M. *Anim. Behav.* **47**, 1333–1341 (1994).
3. Domb, L. G. *Sexual Swellings in Wild Baboons* (Papio cynocephalus anubis) at Gombe National Park, Tanzania. Thesis, Harvard Univ. (2000).
4. Fa, J. E. & Southwick, C. H. (eds) *Ecology and Behavior of Food-Enhanced Primate Groups* (Liss, New York, 1988).
5. Clutton-Brock, T. H. & Harvey, P. H. *J. Zool., Lond.* **183**, 1–39 (1977).
6. Packer, C. *Anim. Behav.* **27**, 1–36 (1979).
7. Alberts, S. & Altmann, J. *Am. Nat.* **145**, 279–306 (1995).
8. Altmann, J. in *Primate Males: Causes and Consequences of Variation in Group Composition* (ed. Kappeler, P. M.) 236–247 (Cambridge Univ. Press, Cambridge, 2000).
9. Nunn, C. L. *Anim. Behav.* **58**, 229–246 (1999).
10. Hamilton, W. J. in *Primate Paternalism* (ed. Taub, D. M.) 309–335 (Van Nostrand Reinhold, New York, 1984).

Domb and Pagel reply — Zinner *et al.* question our finding that the size of a female wild baboon's sexual swellings predicts her lifetime reproductive success, suggesting that we should have controlled for female height in our analyses of female fitness. But the issue is not whether a female's fitness is independent of her height, but whether males draw inferences about her fitness from her sexual swelling, and do so independently of her height. Our results indicate that they do^{1,2}, supporting the hypothesis that sexual swellings advertise female reproductive value.

There is no record to suggest that male baboons, or male primates of any species, prefer taller females. We have shown that males base their costly mating behaviours on the female's swelling size and not on female height, age or social rank. The amount of aggression from other males that a male must tolerate to consort a female (the key male behavioural trait³) is positively correlated with the size of her swelling (swelling length), statistically controlling for the effects of female height, age and rank ($r = 0.59$, $n = 13$). By comparison, males do not receive more aggression for consorting taller females ($r = 0.41$, $n = 13$; $r = -0.069$, $n = 13$, controlling for swelling size; $r = -0.25$ controlling for swelling size, female age and female family rank). Contrary to the claim by Zinner *et al.*, we did control for troop differences in these analyses^{1,2}, and in a way that normalized the variances, unlike their approach, which neglects this statistical assumption.

The suggestion of Zinner *et al.* that we should have analysed male interest in females' swellings when more than one female was in oestrus is based on a misinterpretation of our results. We followed an individual focal female, and not all oestrous females simultaneously. Focal females were

not selected according to their swelling size. If two females were simultaneously in their maximally swollen phase, male interest towards the focal female at any given period would be influenced by the presence of the other female; if the other female's swelling was larger than that of the focal female, then male interest in the focal female could be very low.

From our results, this could be incorrectly inferred as failing to support our hypothesis. The only way to untangle such effects is to investigate all possible pairs (triples and so on) of females, following males and females simultaneously. We therefore limited our tests to those situations in which a single female was in oestrus. This procedure automatically controls for other effects and tests the hypothesis.

Zinner *et al.* repeat an earlier idea^{3,4} that swellings are probabilistic signals of the timing of ovulation, which has since fallen out of favour^{5–7}. Nunn has pointed out⁸ that females in species with sexual swellings are attractive and mate for many more days than females in species without sexual swellings, casting further doubt on whether swellings accurately signal ovulation — in fact, they may function to confuse the timing of ovulation^{9,10}. We tested the second hypothesis of Zinner *et al.* but have found no support for it². This leaves the third hypothesis (ours): that swelling size signals enduring differences between females in their ability to conceive and raise offspring.

The contention of Zinner *et al.* that the reliable-indicator hypothesis is unsupported is therefore based on a misunderstanding of this hypothesis and of our results. Males endure higher costs to mate with females with larger sexual swellings², but not with those that are taller¹. Females with larger sexual swellings also have higher lifetime reproductive success². Baboons provide little or no paternal care, and male–male aggression over females is costly. It is important for males to direct their mating efforts towards the females of the highest fitness, and male selectivity, in turn, makes it advantageous for females to advertise their quality. The size of the sexual swelling seems to do just this.

Leah G. Domb*, Mark Pagel†

*Lawrenceville School, PO Box 6008, Lawrenceville, New Jersey 08648, USA

e-mail: ldomb@lawrenceville.org

†School of Animal and Microbial Sciences, University of Reading, Reading RG6 6AJ, UK

1. Domb, L. G. *Sexual Swellings in Wild Baboons* (Papio cynocephalus anubis) at Gombe National Park, Tanzania. Thesis, Harvard Univ. (2000).
2. Domb, L. G. & Pagel, M. *Nature* **410**, 204–206 (2001).
3. Hausfater, G. *Dominance and Reproduction in Baboons* (Karger, Basel, 1975).
4. Hamilton, W. J. in *Primate Paternalism* (ed. Taub, D. M.) 309–335 (Van Nostrand Reinhold, New York, 1984).
5. Thompson, J. *et al. Biol. Reprod.* **46**, 879–884 (1992).
6. Heistermann, M., Mohle, U., Vervaecke, H., van Elsacker, L. & Hodges, J. *Biol. Reprod.* **55**, 844–853 (1996).

7. Whitten, P. & Russell, E. *Am. J. Primatol.* **40**, 67–82 (1996).
8. Nunn, C. L. *Anim. Behav.* **58**, 229–246 (1999).
9. Boesch, C. & Boesch-Achermann, H. *The Chimpanzees of the Tai Forest: Behavioral Ecology and Evolution* (Oxford Univ. Press, New York, 2000).
10. Wrangham, R. in *Behavioral Diversity in Chimpanzees and Bonobos* (eds Boesch, C., Hohmann, G. & Marchant, L.) (Cambridge Univ. Press, Cambridge, 2002).

COMMUNICATIONS ARISING

Magnetic properties

Parasitic ferromagnetism in a hexaboride?

Surprisingly for a compound with no magnetic element, Young *et al.*¹ have observed ferromagnetism in calcium hexaboride (CaB₆) doped with lanthanum (La) — the system has a ferromagnetic Curie temperature as high as 600 K, which is comparable to that of transition-metal ferromagnets such as iron (Fe). Here we show that high-temperature ferromagnetism in this CaB₆ system is not intrinsic but that it is instead due to alien phases of iron and boride.

Our CaB₆ and LaB₆ samples were synthesized in a solid-state reaction² using three types of crucible (BN, ZrO₂ and MgO). We used an acid cleaning procedure to remove magnetic ions, such as iron, from each sample. Samples were immersed in dilute HCl and the amount of soluble iron was measured using an inductively coupled plasma instrument (Varian). The magnetization of each sample was determined before and after acid treatment by using a magnetic balance magnetometer (Cahn).

Figure 1a shows the temperature dependence of the magnetization $M(T)$ of CaB₆ synthesized using a BN crucible. Before acid treatment, a ferromagnetic feature is evident at around 600 K and 1,000 K; however, this disappears after acid treatment. This disappearance may be due either to an inherent ferromagnetism in CaB₆, with the surface state being altered by acid treatment, or to the removal from CaB₆ of ferromagnetic atoms such as iron (M. Sato *et al.*, unpublished results) by the acid cleaning procedure. As iron is detected in the acid, the second possibility is more probable.

Figure 1b shows the corresponding $M(T)$ curves measured for LaB₆. The sample that was synthesized using a BN crucible exhibits diamagnetic and temperature-independent behaviour. However, the sample synthesized using a ZrO₂ crucible shows ferromagnetic behaviour similar to that evident in Fig. 1a. The Curie temperature of LaB₆ coincides with that of CaB₆. Again, acid treatment reduces the ferromagnetic magnetization (Fig. 1b, middle curve). These results indicate that the ferromagnetism is not inherent to CaB₆ but is

due to a magnetic phase that exists in both CaB_6 and LaB_6 .

We found that there was a strong correlation between ΔM and m_{Fe} , where ΔM is the difference in magnetization at around 90 K (measured before and after acid treatment) and m_{Fe} is the mass of iron dissolved in the acid. There is also a correlation between $\Delta M/m_s$ and m_{Fe}/m_s , where m_s is the sample mass (Fig. 1c). The data correspond well to the saturation magnetization, straight-line curves for the ferromagnets FeB and Fe_2B that have critical temperatures at around 598 K and 1,015 K, respectively³. We conclude that the high-

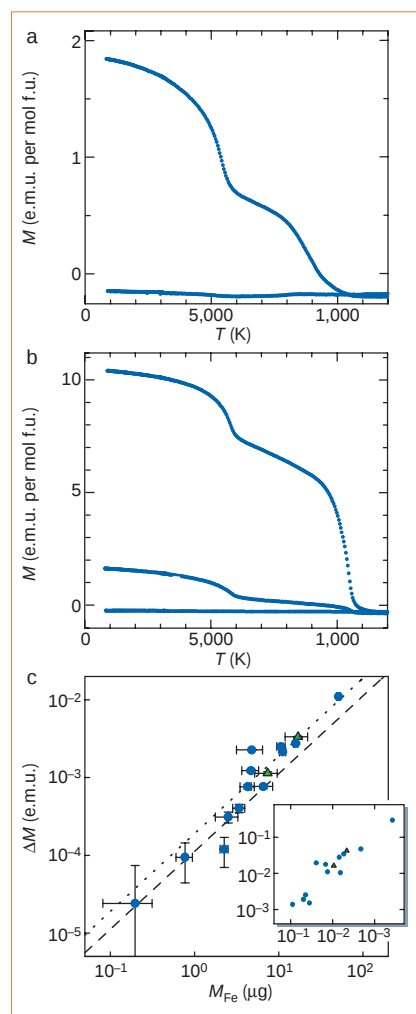


Figure 1 Magnetization of CaB_6 and LaB_6 . **a**, Temperature (T) dependence of magnetization (M , in e.m.u. per mol formula unit) of CaB_6 , measured at 5.1 kOe with decreasing temperature. Top and bottom curves, before and after acid treatment, respectively. **b**, Temperature dependence of magnetization of LaB_6 . Top and middle curves represent the sample synthesized using ZrO_2 before and after acid treatment, respectively; bottom curve represents the sample synthesized using BN. **c**, Correlation between magnetization reduction, ΔM , and iron mass, m_{Fe} . The data are consistent with the curves indicating the saturation magnetization of FeB (dashed line) and Fe_2B (dotted line). Inset, plot of $\Delta M/m_s$ in e.m.u. per gram (m_s , sample mass) against m_{Fe}/m_s in $\mu\text{g g}^{-1}$. Circles, CaB_6 ; triangles, LaB_6 .

temperature ferromagnetism observed by Young *et al.*¹ should be ascribed to these Fe–B phases.

The magnetization of CaB_6 also depends on the type of crucible used for synthesis: M measured at about 5 kOe at room temperature is roughly 0–2, 5–15 or 30 electromagnetic units (e.m.u.) per mol for samples synthesized using the BN, ZrO_2 or MgO crucibles, respectively. Inductively coupled plasma analysis of the used crucibles gave iron weight-concentrations, C_{Fe} , of 2–4, 17–46 and 160 parts per million for the BN, ZrO_2 and MgO crucibles, respectively. We find a correlation between M and C_{Fe} , indicating that the iron atoms moved from the crucible to the sample during synthesis.

Non-doped CaB_6 has also been claimed to exhibit ferromagnetism^{2,4}. It has been suggested⁵ that, as magnetic moments are confined within the sample surface of single-crystalline materials⁶, there can be no intrinsic ferromagnetic moments inside the sample. It is therefore likely that the origin of the ferromagnetism in $\text{Ca}_{1-x}\text{La}_x\text{B}_6$ is the same as in CaB_6 . A sintered material is useful for revealing the origin of the ferromagnetism, as it has a large surface/volume ratio compared with a single crystal. We contend that ours is a reasonable explanation for why almost all samples should exhibit the same Curie temperature, despite their experimental differences.

K. Matsubayashi, M. Maki, T. Tsuzuki, T. Nishioka, N. K. Sato

Department of Physics, Graduate School of Science, Nagoya University, Nagoya 464-8602, Japan
e-mail: kensho@edu3.phys.nagoya-u.ac.jp

1. Young, D. P. *et al.* *Nature* **397**, 412–414 (1999).
2. Morioka, T., Nishioka, T. & Sato, N. K. *J. Phys. Soc. Jpn* **70**, 341–344 (2001).
3. Chikazumi, S. *et al.* (eds) *Handbook of Magnetic Materials* (Asakawa-shoten, Tokyo, 1975).
4. Vonlanthen, P. *et al.* *Phys. Rev. B* **62**, 10076–10082 (2000).
5. Taniguchi, K. *et al.* *Phys. Rev. B* **66**, 064407–1–7 (2002).
6. Kunii, S. *J. Phys. Soc. Jpn* **69**, 3789–3791 (2000).

Young *et al.* reply — The results of Matsubayashi *et al.* on nominally stoichiometric, CaB_6 sintered powders arrive at a conclusion concerning our study¹ on ferromagnetism in this and related materials that differs in several ways from the conclusions of ongoing studies of aluminium-flux-grown single crystals (Z.F., S. Nakatsuji, F. Drymiotis, M. Bizimis, S. Yeo, J.D.T., A. Bianchi and H.R.O., unpublished results).

At issue is whether alien Fe–B phases are responsible for all of the observed ferromagnetism. Our conclusion is partly contained in ref. 2, in which it is stated that strongly interacting defects in off-stoichiometric CaB_6 (that is, comprising a few-tenths of a per cent) that carry magnetic

moments are responsible for the observed ferromagnetic properties. These defects may be iron atoms scavenged during the growth of CaB_6 crystals from boron-rich flux growths. We outline experimentally based arguments (Z.F. *et al.*, unpublished results) against the suggestion of Matsubayashi *et al.* that extrinsic Fe–B phases are the source of the ferromagnetism here. Almost all studies show the surface moments to be removed in acid solution, and these will not be discussed further.

In single crystals of CaB_6 grown from molten aluminium flux (starting composition for Ca:B, 1:1–1:9) with added iron (Ca:Fe, 1:0.01–1:1), high-temperature ferromagnetism is generally observed in crystals with Ca:B ratios greater than 1:6; the magnitude of the ordered moment is 3–5 e.m.u. per mol CaB_6 for all iron concentrations in the flux (Z.F. *et al.*, unpublished results). This lack of dependence of the measured ordered moment on the iron concentration in the flux suggests that alien Fe–B phases are not the source of ferromagnetism. Also, ferromagnetic CaB_6 crystals have a metallic electrical-resistivity characteristic ($d\rho/dT > 0$; ref. 2) that is consistent with the introduction of defects into the material.

If cobalt or nickel is added to the flux in concentrations equal to those of added iron, the crystals of CaB_6 show small-moment ferromagnetism only below about 10 K (Z.F. *et al.*, unpublished results), indicating that cobalt or nickel enters CaB_6 as low-concentration defects in preference to iron, and that cobalt and nickel do not magnetically interact as strongly as iron defects.

A picture emerges in which iron is involved in the weak high-temperature ferromagnetism of CaB_6 and in which off-stoichiometric growth of CaB_6 from excess boron melting results in the scavenging of iron from the flux, giving rise to crystals of CaB_6 containing about 0.1 atomic per cent Fe. Theoretical³ and experimental^{4,5} studies of CaB_6 indicate that the stoichiometric material is a semiconductor. This suggests that the physics here shares many similarities with that of manganese-doped GaAs. It will be interesting to determine the chemistry of the iron defects, as well as how they interact so strongly and conspire to magnetic order at higher temperatures.

D. P. Young, Z. Fisk*, J. D. Thompson, H. R. Ott, S. B. Oseroff, R. G. Goodrich

*National High Magnetic Field Laboratory, Florida State University, Tallahassee, Florida 32310, and Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA

e-mail: fisk@magnet.fsu.edu

1. Young, D. P. *et al.* *Nature* **397**, 412–414 (1999).
2. Fisk, Z., Ott, H. R., Barzykin, V. & Gor'kov, L. P. *Physica B* **312–313**, 808–810 (2002).
3. Tromp, H. J., van Gelderen, P., Kelly, P. J., Brooks, G. & Bobbert, P. A. *Phys. Rev. Lett.* **87**, 16401/1–4 (2001).
4. Vonlanthen, P. *et al.* *Phys. Rev. B* **62**, 10076–10082 (2000).
5. Denlinger, J. D. *et al.* *Phys. Rev. Lett.* **89**, 157601/1–4 (2002).

Serial deletions and duplications suggest a mechanism for the collinearity of *Hoxd* genes in limbs

Marie Kmita*, Nadine Fraudeau*, Yann Hérault*† & Denis Duboule*

* Department of Zoology and Animal Biology, NCCR Frontiers in Genetics, University of Geneva, Sciences III, Quai Ernest Ansermet 30, 1211 Geneva 4, Switzerland

***Hox* genes, located at one end of the *HoxD* cluster, are essential for the development of the extremities of our limbs; that is, the digits. This ‘collinear’ correspondence is accompanied by a gradual decrease in the transcriptional efficiency of the genes. To decipher the underlying regulatory mechanisms, and thus to understand better how digits develop, we engineered a series of deletions and duplications *in vivo*. We find that *HoxD* genes compete for a remote enhancer that recognizes the locus in a polar fashion, with a preference for the 5′ extremity. Modifications in either the number or topography of *Hoxd* loci induced regulatory reallocations affecting both the number and morphology of digits. These results demonstrate why genes located at the extremity of the cluster are expressed at the distal end of the limbs, following a gradual reduction in transcriptional efficiency, and thus highlight the mechanistic nature of collinearity in limbs.**

During vertebrate evolution, *Hox* genes were co-opted to achieve a variety of functions in addition to their ancestral role in organizing structures along the trunk axis. One example is the limb skeleton, which is patterned by genes from the *HoxA* and *HoxD* clusters^{1,2}. These transcription factors are essential for limb development^{3,4}, and their collinear regulation is similar to that observed in the trunk: genes located in the middle of the *HoxD* complex (for example, *Hoxd9*) are expressed in proximal areas of the limb bud whereas genes that are located in a more 5′ direction (for example, *Hoxd11*) have a more distal expression, in the future forearm region. Accordingly, the most distal parts of the limb, the future hands and feet, express the four genes situated in the most 5′ location (*Hoxd10* to *Hoxd13*). Therefore, the skeletal organization of the limb is prefigured in the genomic topology of these genes^{5–7} (Fig. 1a). Furthermore, a decrease in transcription efficiency correlates with gene order (quantitative collinearity⁸; Fig. 1a).

The proper regulation and function of these genes is critical, as their inactivation leads to the absence of the corresponding structures: the loss of *Hoxd11* and *Hoxa11* function is accompanied by the loss of forearms³, whereas the absence of both *Hoxd13* and *Hoxa13* prevents the development of hands and feet^{9–11}. These phenotypes were difficult to evaluate, owing to redundant and compensatory mechanisms, and to a functional hierarchy attributing a prevalent role to the most ‘posterior’ proteins over more ‘anterior’ ones¹². We reported previously that the element necessary for digit expression of the four *Hoxd* genes situated in the most 5′ location (*Hoxd13* to *Hoxd10*), and of the neighbouring gene *Evx2*, is located outside of the *HoxD* cluster^{13–15}. Moreover, other *Hox*^{16,17} and foreign¹⁸ promoters responded to this enhancer whenever introduced at the extremity of the locus, and this response was independent of their transcriptional orientation. This raised the question of the specificity of action of this regulatory element; that is, its potential to discriminate between various, distantly located *Hox* promoters and to lead reproducibly to the observed collinear expression patterns, thus producing the correct digit morphology.

To dissect the function of these genes and to investigate the mechanism underlying collinearity in limbs, we carried out a systematic deletion and duplication approach of *Hoxd* loci, reason-

ing that modifications in the presence, the absence, the position or the number of genes would provide insights into both aspects. We used targeted meiotic recombination (TAMERE¹⁹) to produce unequal recombination between the *Hoxd13*, *Hoxd12* and *Hoxd11* loci. Furthermore, some deletions and duplications were engineered along with other mutations in *cis* (Fig. 1b). This allelic series clarified functional issues and provided explanations as to why genes situated in 5′ of the *HoxD* cluster preferentially respond to this regulation, and why they do it with a decrease in transcriptional efficiency; thereby deciphering part of the mechanism underlying collinearity in limbs.

Serial deletions

We generated deletions and duplications by crossing mice containing *loxP* sites between *Evx2* and *Hoxd13*, between *Hoxd13* and *Hoxd12*, between *Hoxd12* and *Hoxd11*, or between *Hoxd11* and *Hoxd10* (Fig. 1b). After recombination, all remaining genes were fully functional. This series of deficiencies and duplications involved from one to three loci, such as deletion (or duplication) of *Hoxd13*, of *Hoxd13* and *Hoxd12* in *cis*, or of three loci (Fig. 1b). The phenotypes of the homozygous animals were examined. Severe skeletal phenotypes were scored only on excision of *Hoxd13*—deletion of both *Hoxd12* and *Hoxd11* did not lead to important alterations (not shown), in agreement with the functional dominance of *Hoxd13* over other posterior *Hoxd* genes²⁰. However, although deletions containing *Hoxd13* impinged on digit morphology, the alterations appeared unexpectedly different from those described for the loss of function of the corresponding genes generated by disruption of coding sequences.

This finding was exemplified by deleting the *Hoxd13* locus, which generated a mild digit alteration in comparison with animals carrying insertional mutations^{21–23} (Fig. 2a, b) that show a reduction in digit length, a supernumerary posterior digit and an overall ill-formed and stiffer aspect of bony elements (Fig. 2a). In contrast, after deletion of *Hoxd13*, digit length was no longer affected, apart from a slight reduction of digits II and V, and no supernumerary digit was scored. Altogether, the morphology was close to that of control animals (Fig. 2b, c). Furthermore, mice carrying the deletion had normal hindlimbs, without the skeletal defects associated with *Hoxd13* loss of function²¹. Consequently, whereas *Hoxd13*

† Present address: Molecular and Experimental Genetics, FRE2358, CNRS, Institut de Transgénèse, rue de la Férollerie, 3B, 45071, Orléans cedex 2, France.

disruptions induce severe limb alterations, the corresponding deletion has little effect on these structures. This phenomenon illustrates the difficulty over assigning an objective meaning to the concept of 'gene function', as different full loss-of-function alleles have distinct consequences on the final morphology. Within such gene clusters, the 'function' of a given gene cannot be considered on its own, but instead must be integrated into a larger functional context, that of the cluster itself.

We analysed the expression of neighbouring genes to the disrupted allele and found no difference with control (Fig. 2). In contrast, deleting the *Hoxd13* locus resulted in a deregulation of the 3' neighbouring gene (*Hoxd12*; Fig. 2a, b). A robust enhancement in *Hoxd12* transcription was observed, and *Hoxd12* transcripts were detected in the most anterior aspect of the distal bud, within the presumptive domain of digit I, where only *Hoxd13* transcripts were seen in control animals²⁴ (Fig. 2b, c). This gain of expression explained why these mice were less affected than mice with disrupted alleles: increased transcription of *Hoxd12* partially substituted for the absence of *Hoxd13*.

Therefore, moving *Hoxd12* to the position of *Hoxd13*, with respect to the extremity of the cluster, led to a regulatory reallocation of *Hoxd12*, which became transcribed like *Hoxd13* (Fig. 2b, c).

To see whether this was specific for *Hoxd12* or if it reflected a general versatility of *Hoxd* promoters, we analysed a deletion whereby *Hoxd11* was placed at the position of *Hoxd13*, and compared it with a double disrupted allele, inactivating both *Hoxd13* and *Hoxd12* (ref. 25) (Fig. 1b). Again, limbs of deleted mice appeared different from those of the double inactivation in *cis*²⁵. In this case, animals carrying the two-gene deletion were affected more severely than with the double inactivation in *cis* (Fig. 2e, f), and up to ten supernumerary cartilaginous condensations were scored and the entire autopod (the most distal skeletal limb segment) was markedly ill formed (Fig. 2). Metacarpal bones appeared chaotic, with partial fusions, and phalanges failed to elongate properly.

This severe phenotype correlated with both an enhancement of *Hoxd11* transcription and the anterior expansion of its transcript domain (Fig. 2f). Again, *Hoxd11* behaved like *Hoxd13* when located at the (relative) *Hoxd13* position. In contrast, *Hoxd11* expression was normal in the double inactivation mutant (Fig. 2d, e), indicating that the severity of the defects was associated with the upregulation of *Hoxd11*. The potential of *Hoxd11* to trigger supernumerary chondrogenic condensations was verified by using a deletion removing *Hoxd13*, *Hoxd12* and *Hoxd11*; in this condition, the number of digits was back to five and *Hoxd10* expression was

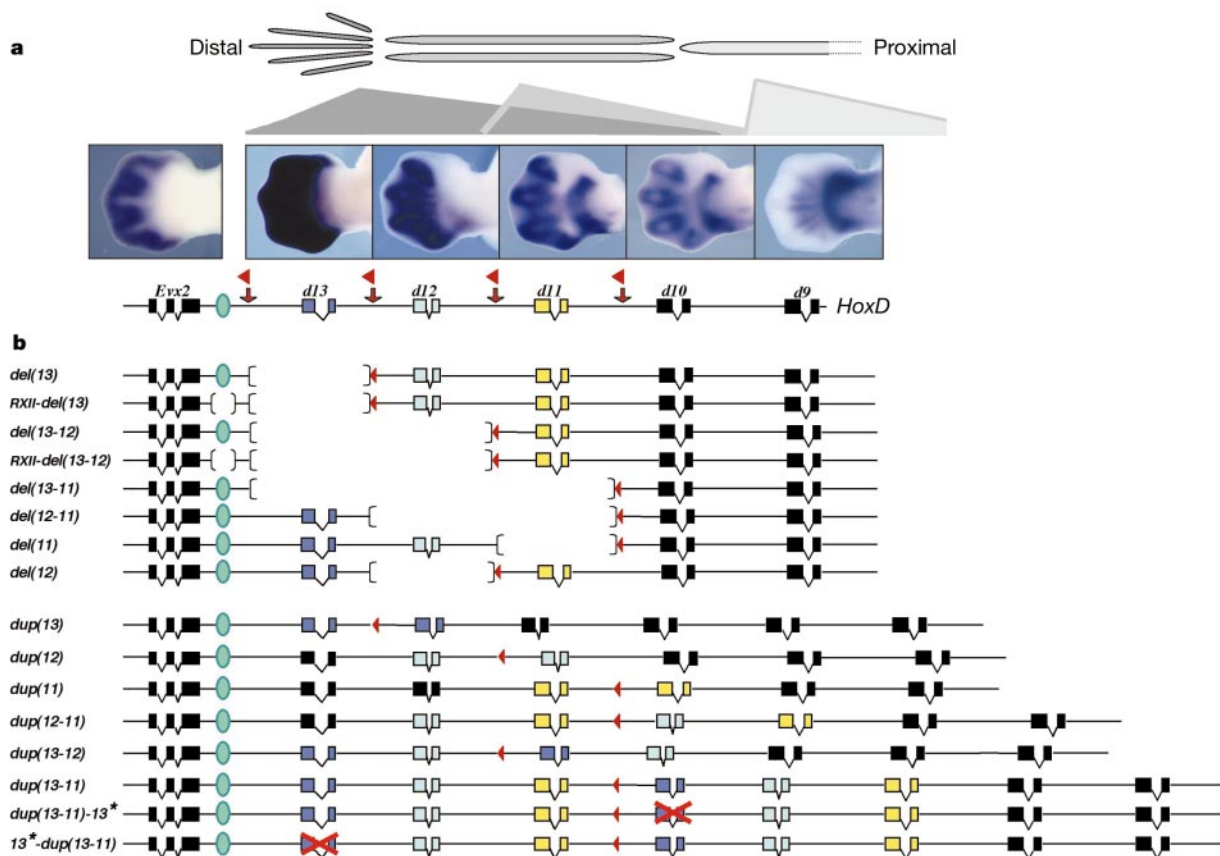


Figure 1 Collinearity in developing limbs. **a**, Scheme of a forelimb skeleton (top) and expression of five of the most 5' *Hoxd* genes in limb buds at day 12.5 (bottom). Following their genomic order, expression of these genes is progressively restricted to the most distal part of the developing limbs (grey triangles). *Hoxd* genes are also expressed following a gradient of transcriptional efficiency within the distal domain itself, such that *Hoxd13* is most strongly expressed, whereas genes located in a more 3' direction display progressively less robust expression levels (quantitative collinearity⁸). Whereas only *Hoxd13* is expressed in presumptive digit I, *Hoxd9* is barely detectable in any of the digits. **b**, Allelic series. The complete set of deletions and reciprocal duplications involving one, two or three loci within the posterior *Hoxd* complex (from *Hoxd13* to *Hoxd11*) are shown. Parental alleles (see Methods) were selected as a function of the location of *loxP* sites (red

triangles) in between transcription units (red arrows). Brackets indicate the deletion breakpoints; *Hoxd13* (dark blue), *Hoxd12* (light blue) and *Hoxd11* (yellow) are shown whenever not deleted. The deletion of *Hoxd13* alone, as well as that involving both *Hoxd13* and *Hoxd12*, was produced with and without an additional deletion in *cis* of a 1.2-kb DNA fragment containing region XII (*del(13)*, *del(13-12)*, and *RXII-del(13)*, *RXII-del(13-12)*, respectively). The location of region XII is indicated (green oval). The duplication alleles are shown with the same colour code (bottom). The position of the remaining *loxP* site is indicated (red triangles). For the duplication of the three loci (*dup(13-11)*), two additional alleles were produced by using loss-of-function alleles of *Hoxd13* as parental stocks, such that one of the two copies of *Hoxd13* was inactivated (red cross) either in the most 5' location (*13*-dup(13-11)*) or in the 3' location (*dup(13-11)-13**).

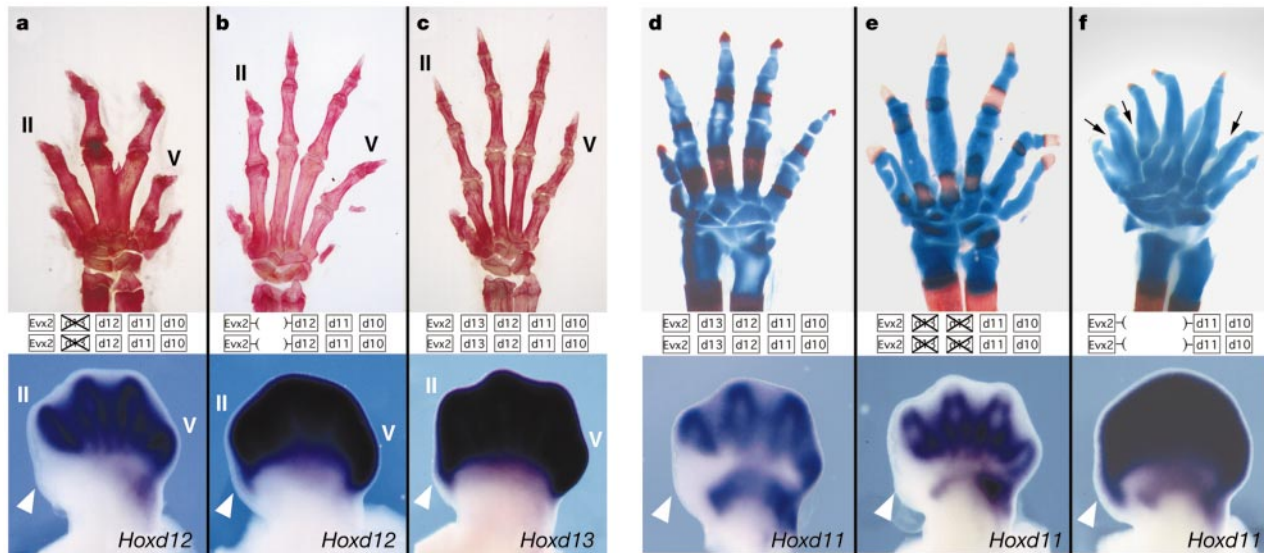


Figure 2 Targeted deletions induce regulatory reallocations. Comparison between digit phenotypes (top) and expression of 5' *Hoxd* genes (bottom) associated with either the disruption or the deletion of the corresponding loci. Crosses indicate gene inactivation; brackets indicate deletion breakpoints. **a**, Inactivation of *Hoxd13* leads to an overall reduction in the size of digits, partial fusion between digits III and IV, and a supernumerary digit in most cases²³. **b**, In contrast, deletion of the same locus has little effect (compare with control in **c**). In this latter case, a strong gain of expression of *Hoxd12* was scored,

resembling the *Hoxd13* expression pattern (**b**, **c**). **f**, Deletion of both *Hoxd13* and *Hoxd12* loci induce a severe polydactyly, and fusions (arrows) indicate an even larger number of pre-chondrogenic condensations. **e**, This phenotype is not observed in the corresponding double inactivation allele. In this case, a robust gain of *Hoxd11* expression is associated with deletion of both *Hoxd13* and *Hoxd12* loci (**d**, **f**), whereas it is absent from the disruption of both loci (**e**). In both gains of expression (**b**, **f**), the profiles are reminiscent of wild-type *Hoxd13* (**c**), including in presumptive digit I (white arrowheads).

re-enforced in digits (see Supplementary Information 1).

This link between *Hoxd11* and polydactyly indicated that supernumerary digits previously associated with the loss of *Hoxd13* function^{21,25} were in fact induced by the release of a functional suppression exerted by *Hoxd13* over *Hoxd11* (refs 16, 26). The phenotype of mice in which both *Hoxd12* and *Hoxd11* expressions were strongly enhanced (see below) indicated that *Hoxd12*, as *Hoxd13*, was able to suppress the gain-of-function effect of

Hoxd11; although these mice had variable phenotypes, polydactyly was never scored despite high *Hoxd11* expression in digits and the absence of *Hoxd13* product. These results show that *Hoxd12*, similar to *Hoxd13*, exerts posterior prevalence; that is, it can functionally abrogate the effects of other *Hoxd* proteins during development.

Regulatory reallocations

Such gains of expression were also scored in heterozygous embryos, suggesting that interactions in *cis* between a remote enhancer and various *Hoxd* promoters had been modified after the deletion of promoter regions. Removing the *Hoxd13* locus, for instance, would redirect the enhancer towards the *Hoxd12* promoter with a comparable efficiency. Similarly, a double deletion would give *Hoxd11* the leading position to respond to the enhancer in the absence of two formerly competing promoters. Notably, only the nearest neighbouring gene was transcriptionally enhanced (Fig. 2), whereas genes located further apart, such as *Hoxd11* in the deletion of *Hoxd13*, were hardly affected, indicating a preference of the enhancer sequence for whichever promoter was located at the 5' extremity of the cluster. This enhancer tropism for one extremity of the cluster was further confirmed by deletions of *Hoxd12* and *Hoxd11*, either alone or combined (Fig. 1); although deletion of one locus did not change *Hoxd* gene expression, deletion of both loci in *cis* merely induced a slight upregulation of *Hoxd10* (not shown).

These observations suggest that *Hoxd13* normally competes more efficiently for interaction with the digit enhancer, as seen in anterior distal cells where it is the only *Hoxd* gene transcribed. In the β -globin cluster, proximity to the locus control region modulates the relative expression of the genes during development by altering expression through promoter competition²⁷. This parallel between *HoxD* and β -globin loci suggests that promoter competition may be a generic mechanism whereby differential transcriptional regulation is generated within sets of target genes controlled by common enhancers. Here, because *Hoxd13* is heavily transcribed in a wild-type context, potential variations in our deletions are difficult to assess. It is nevertheless clear that, in the presence of *Hoxd13* such regulatory reallocations are minor. As the gene at the extremity of the cluster is

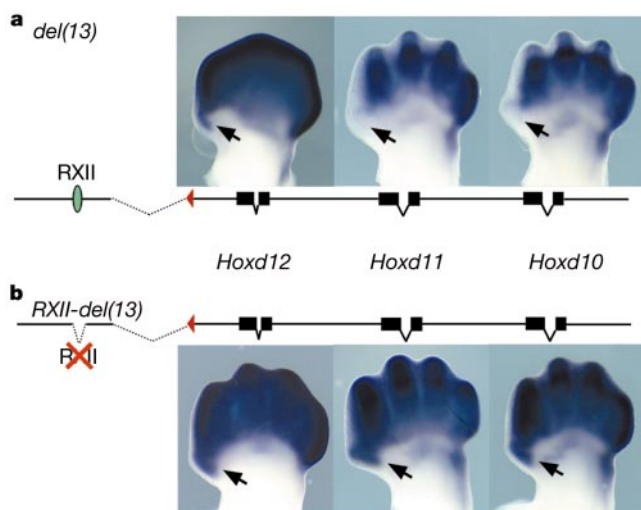


Figure 3 The evolutionarily conserved region RXII cooperates with the *Hoxd13* locus for positioning the enhancer in the 5' end of the cluster. **a**, Expression of three neighbouring *Hoxd* genes in the limbs of 13.5-day-old embryos carrying a deletion of the *Hoxd13* locus. Although there is a gain of *Hoxd12* expression in the entire distal domain, both *Hoxd11* and *Hoxd10* transcription remains virtually unchanged. **b**, In contrast, further deletion of region XII, in the absence of the *Hoxd13* locus, induces a gain of expression for all three genes, including in presumptive digit I (black arrows), thus abrogating quantitative collinearity.

always the most efficiently transcribed, we looked for the cause of this preferential interaction between the enhancer and this part of the cluster.

Enhancer tropism

DNA sequence comparisons between vertebrate genomes revealed two stretches of high conservation around *Hoxd13*: regions XI (ref. 28) and XII (ref. 29) (RXI and RXII, respectively). Whereas RXI was excised in all deletions containing the *Hoxd13* locus, RXII, located between *Hoxd13* and *Evx2*, was always left in place. To investigate whether RXII was instrumental in targeting the enhancer towards the extremity of the cluster, we produced another set of deletions wherein both RXII and the *Hoxd13* locus were removed (Fig. 1b; *RXII-del(13)* and *RXII-del(13-12)*).

Embryos carrying the *Hoxd13* and RXII deletions showed the expected increase in *Hoxd12* expression; however, they also displayed enhanced transcription of *Hoxd11* and *Hoxd10*, unlike embryos lacking the *Hoxd13* locus alone (Fig. 3). We concluded that removing both RXII and the *Hoxd13* locus induced all of the remaining 5' *Hoxd* genes to be expressed in a similar pattern to *Hoxd13*—a break in quantitative collinearity. The combined deletion of RXII, *Hoxd13* and *Hoxd12* confirmed this result; however, deletion of RXII from an otherwise intact cluster had no detectable effect on 5' *Hoxd* gene regulation (not shown). Potential

elements within RXII that could help in the targeting of the enhancer remain unknown.

We next wondered whether this RXII-dependent enhancer tropism could account for the weak (if any) expression of *Hoxd9* in digits. *Hoxd9* expression was strong with one copy of the combined *RXII-del(13-12)* deletion, and two copies triggered expression in digit I further demonstrating the principal function of RXII in targeting the enhancer towards the 5' end of the cluster. Therefore, regulatory reallocations also involve a gene that is normally scarcely expressed in presumptive digits (Supplementary Information 2), indicating that the weakness of *Hoxd9* expression in developing digits was primarily due to the presence of competing promoters upstream, rather than to the presence of boundary elements. Together, the tropism of the enhancer for the 5' extremity of the cluster gave an indication of why 5' *Hoxd* genes responded to this regulation during normal development.

Targeted gene duplications

This versatility in enhancer–promoter interactions was further analysed using targeted duplications that involved one, two or three loci (Fig. 1). We first looked at animals carrying two copies of the various duplications, and we did not detect any significant morphological alteration. This was not unexpected, as all duplications contained at least one copy of *Hoxd13*, whose function seems to trigger the developing limb to enter a 'terminal' programme, both by changing the pattern of chondrogenic condensations and by suppressing the function of other proteins.

We analysed the impact of the duplications on flanking genes and found no critical difference in expression whenever a single locus was duplicated. However, larger duplications led to significant modifications. For example, both *Hoxd10* and *Hoxd11* were downregulated in limb buds of embryos duplicated for both *Hoxd13* and *Hoxd12* (Fig. 4a), an effect restricted to presumptive digits—staining in other parts of these embryos was indistinguishable from controls. This suggested that supernumerary loci efficiently out-competed promoters in a more 3' location for the activity of the enhancer, a conclusion re-enforced by the largest duplication, which involved a duplication from *Hoxd13* to *Hoxd11* (Fig. 1). With three supernumerary loci in a 5' location, expression of *Hoxd10* was severely affected: restricted to digits III and IV (Fig. 4a). A correlation was thus observed between the extent of *Hoxd10* downregulation and the number of genes added between *Hoxd10* and the digit enhancer. Notably, the double *Hoxd12 Hoxd11* and *Hoxd13 Hoxd12* duplications had distinct effects on *Hoxd10*, and a significantly weaker impact was obtained with the former (not shown), even though the size of added DNA was comparable. Therefore, although both the number of promoters and the size of the DNA inserted were the same, the *Hoxd13 Hoxd12* supernumerary promoters could titrate out the activity of the enhancer sequence more efficiently than *Hoxd12 Hoxd11*, further supporting a target-dependent process rather than a mere distance effect.

To investigate this conclusion we compared the response of the same locus when simultaneously localized at different positions. We produced two new duplications of the *Hoxd13*, *Hoxd12* and *Hoxd11* loci in which either one of the *Hoxd13* copies was inactivated (Fig. 1b). The two lines of mice were identical in all respects, yet they differed in the position of the inactivated *Hoxd13* copy: one line carried the mutated copy at the extremity of the complex, whereas the other carried the mutated loci as the fourth transcription unit from the 5' end (Fig. 4b). Animals of this latter genotype had limbs indistinguishable from controls. In contrast, inactivation of the copy at the 5' end generated digit defects indicative of a strong *Hoxd13* deficit (Fig. 4b), demonstrating that the non-inactivated *Hoxd13* copy could not compensate for the loss of function of the copy located at the 5' end. In fact, the phenotype of mice of this line indicated that this functional copy was expressed at a low level. This experiment demonstrates that in the presence of two identical

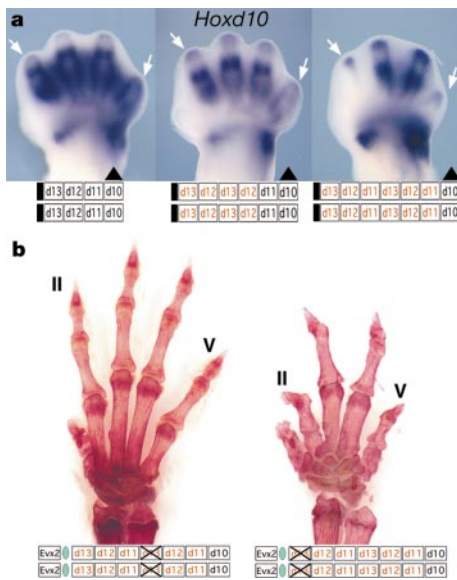


Figure 4 Supernumerary loci titrate the effect of the digit enhancer in a polar fashion. **a**, *Hoxd10* expression in a control limb (left), a limb with a two-gene duplication (middle), or a three-gene duplication (right). Black rectangles are the 5' extremities of the cluster and duplicated genes are in red. Insertion of extra loci upstream of *Hoxd10* induce its downregulation in distal limbs. The downregulation of *Hoxd10* transcription, after duplication of both *Hoxd13* and *Hoxd12* (middle), is accentuated by the additional duplication of the *Hoxd11* locus (right), suggesting a stepwise enhancer titration along with the number of supernumerary promoters. **b**, The digit enhancer works through a polar mechanism. Digits of mice carrying the same three-gene duplication, but with either one of the *Hoxd13* copies inactivated (black cross), are shown. These mice have identical genetic configurations, except for mutated *Hoxd13*, which is located at different positions in the two stocks. The phenotypes are distinct. Inactivation of the copy located in the most 5' direction produces severely affected hands (right), despite the presence of a wild-type copy of *Hoxd13*, whereas animals carrying the mutated copy in the middle of the duplication (left) have normal digit skeletons, indicating that the enhancer sequence is mostly concerned with the copy located in the 5' position, at the expense of the other one. In these alleles, region XII (green oval) was not part of the duplicated DNA fragment, and thus remained at the 5' end of the cluster.

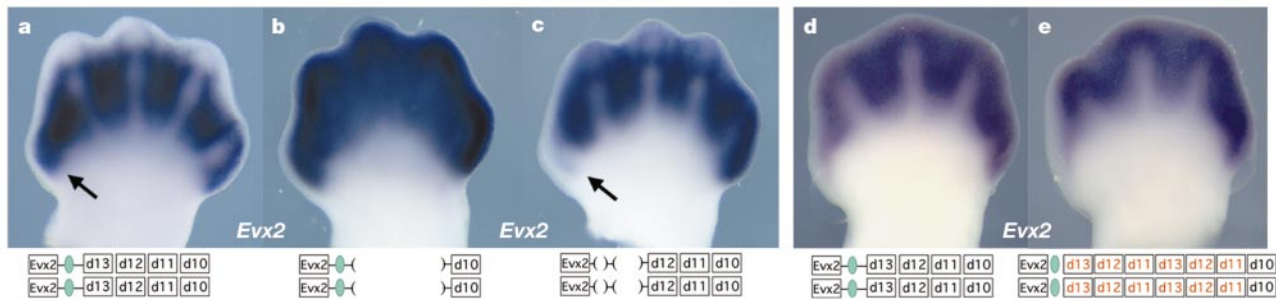


Figure 5 Impact of *Hoxd* loci deletions and duplications on the regulation of *Evx2*. **a**, Normal *Evx2* expression profile. **b**, Deletion of three genes induces a gain of expression, unlike deletions restricted to one or two genes. **c**, Deletion of both RXII and the *Hoxd13* locus, if anything, induces a downregulation of *Evx2*, in particular in presumptive digit I,

whereas 5' *Hoxd* genes are upregulated (see Fig. 3), suggesting that *Evx2* competes less efficiently for the digit enhancer than *Hox* promoters. **d**, **e**, Duplications have no significant effect on *Evx2*, as expected from its proximity to both RXII (green oval) and the *Hoxd13* locus.

promoter regions, the 5' end locus competes more efficiently for interaction with the digit enhancer, probably helped by region XII whose location at the 5' end of the complex remained unmodified in these alleles.

Evx2 regulation

Because *Evx2* is located just upstream of *Hoxd13*, in a reverse transcriptional orientation, its promoter is neighbouring both that of *Hoxd13* and RXII. As *Evx2* responds to the digit enhancer in a similar way to *Hoxd* genes, we monitored its transcriptional behaviour in this allelic series and found it distinct from that of *Hoxd* genes. For instance, *Evx2* expression was significantly enhanced only in the largest deletion (Fig. 5a, b), showing that regulatory reallocations would primarily concern *Hoxd* promoters rather than an unrelated promoter; indeed, *Evx2* was upregulated only when most 5' *Hoxd* promoters were deleted (Fig. 5c). When deletions were combined with excision of RXII, *Evx2* expression became barely detectable in digit I, where this gene is normally expressed together with *Hoxd13* (Fig. 5a, c; arrows), whereas there was a gain of expression of *Hoxd* genes there (Fig. 3).

This observation suggested that in presumptive digit I cells the *Evx2* promoter was out-competed by other *Hoxd* promoters in the absence of both RXII and the *Hoxd13* locus. When present, these latter two elements targeted the action of the enhancer towards this part of the cluster, leading to an expression of *Evx2* that resembled

that of *Hoxd13*. This proximity effect further emphasizes the sequence-specific tropism of the enhancer. In various duplicated alleles, *Evx2* behaved as expected for a gene located between the duplication and the enhancer. Because in all duplications *Evx2* maintained its position relative to both RXII and *Hoxd13*, its expression was not detectably modified, as exemplified by the triple *Hoxd13* to *Hoxd11* duplication (Fig. 5d, e).

Collinearity in limbs

Enhancer preference for the extremity of the cluster may result from different mechanisms. Preferential 5' transcriptional activation may reflect a distance effect, with the gene located at the most 5' position being closest to the enhancer. A distance effect could be produced either by a stochastic process, or through a DNA scanning mechanism, guiding the enhancer towards the closest extremity of the cluster. Alternatively, a sequence-specific mechanism may help to establish a preferential interaction. Our results favour the second alternative: first, duplications introducing additional DNA of the same length but at different loci had distinct effects on the transcription of the gene located at a more 3' position. Second, RXII (a DNA fragment that displays sequence conservation with the chicken genome²⁹) was required along with the *Hoxd13* locus to implement the position-dependent, preferential activation. Removal of both RXII and the *Hoxd13* locus abrogated quantitative collinearity.

Therefore, although the refinement of 5' *Hoxd* gene expression in digits relies on promoter competition, as for the β -globin locus, the mechanisms underlying this competition are probably different in the two systems: in the β -globin locus, the relative distance to the locus control region impacts on preferential activation³⁰. Yet the efficiency of competition can vary in a gene-specific manner, as β and ϵ , at the same position, compete differently with the α gene^{31,32}. We show that competition between *Hoxd* promoters for the activity of the digit enhancer is not primarily related to gene identity and mostly—although not entirely—depends on position, owing to the existence of regulatory sequences that assist the enhancer to find the right target with a high probability.

The experiments described here provide a mechanistic explanation as to how spatial and quantitative collinearity are implemented during limb development. First, a group of genes at the extremity of the cluster respond to a global digit enhancer sequence located upstream. The enhancer contacts this locus as well as other potentially unrelated loci located nearby (Fig. 6a). Once the interaction is initiated, a sequence-specific process might refine this contact by directing and strengthening the interaction towards the 5' extremity of the cluster (Fig. 6b). This mechanism would underlie the correspondence between the extremity of the complex and the distal parts of the limbs. Second, through the same sequence-specific mechanism, the enhancer would display a strong preference

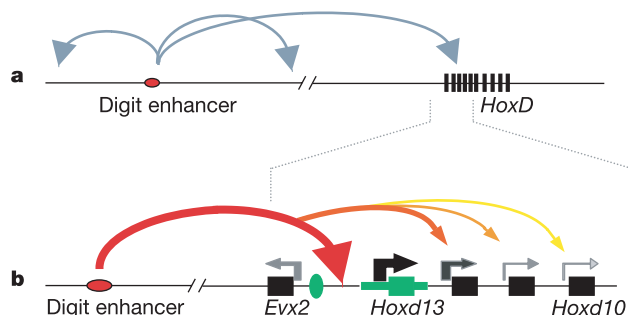


Figure 6 A potential mechanism underlying collinearity in limbs. **a**, The digit enhancer (red oval) acts over a long range and can interact with any transcription unit located in its realm of action (blue arrows), including the *HoxD* cluster. **b**, Once a contact has been established with the *Hoxd* complex, a sequence-specific mechanism, driven by RXII (green oval) and another sequence located in the *Hoxd13* locus (green bar), targets most of the enhancer activity to the 5' extremity of the complex (thick red arrow), leading to a preferential activation of the most 5' gene (*Hoxd13*; black arrow). Accordingly, the enhancer becomes gradually less efficient (orange to yellow arrows) in controlling promoters located further apart in the 3' direction (grey arrows), thus accounting for quantitative collinearity.

for the promoter located at the most 5' position of this subset of genes. As a consequence, it would gradually lose transcriptional efficiency with genes located further apart (Fig. 6), thus accounting for both 'quantitative' and 'reverse'²³ collinearity (the reason why *Hoxd13* is expressed more widely than *Hoxd12*, which conflicts with a strict view of collinearity). This could be due to some neighbouring effect—the enhancer is able to activate promoters located nearby concomitantly with its interaction with the *Hoxd13* locus—or it could be a result of a more stochastic process, such as a flip-flop mechanism—the enhancer is directed towards these neighbour promoters with a reduced probability. Our experimental design cannot discriminate between these two alternatives.

In the light of these results, one may wonder whether the biological relevance of this complex mechanism could not be to secure a high level of *Hoxd13* in distal limbs, due to the importance of this gene in making hands and feet. In this model, quantitative collinearity may have evolved as a by-product of this mechanism, and may reflect the looseness of the process. Analyses of more configurations produced by means of our meiotic recombination strategy will clarify this issue. □

Methods

Meiotic recombination

Genomic rearrangements within the *HoxD* complex were produced *in vivo* through our targeted meiotic recombination approach¹⁹. Mice homozygous for a *HoxD* allele containing a *loxP* site at a given position were crossed with mice hemizygous for the *Sycp1/Cre* transgene³³. Their progeny were subsequently crossed with animals homozygous for another allele of the *HoxD* cluster, but containing a *loxP* site at a different position, to generate 'transloxer' animals; that is, male mice with meiotic cells where targeted recombination can occur. After *trans*-allelic recombination between *loxP* sites located at distinct positions within the *HoxD* cluster, unequal chromosomal exchanges gave rise to the deletion (del) and reciprocal duplication (dup) of the DNA fragment located between the two selected *loxP* sites. Transloxer males were mated with wild-type females, such that the translocated alleles that segregated in haploid germ cells were recovered in their progeny. Southern blot genotyping using tail DNA was used to screen for the progeny of each transloxer, and animals heterozygous for a given targeted recombination event were recovered with a frequency from 1 to 10%. The various alleles that we used as starting material to produce transloxer males are provided as in Supplementary Information 3.

Genetics

For some recombinations, one parental allele (*EvD^{Gell}*) also contained a mutation within the *Evx2* gene³⁴. (For the sake of clarity, and because this mutation did not interfere with the phenotypes or with the regulatory effects associated with the resulting alleles, it is not mentioned in Fig. 1.) For each of the newly produced alleles, we generated lines by selection of heterozygous mice that had the *SYCP1/Cre* transgene segregated out. Mice homozygous for a given deletion were obtained directly in the F₂ progeny and identified by Southern blot. However, heterozygous compared with homozygous mutant mice carrying a duplication allele could not be discriminated by the same strategy (except for the duplication of *Hoxd12*), as no unique restriction sites were found flanking the large duplicated DNA fragments (from 14 to 45 kilobases). Therefore, an indirect strategy to select homozygous animals was designed whereby we systematically produced dup/del *trans*-heterozygous mice, which were subsequently used for sibling crosses. In the progeny of these latter crosses, the genotypes obtained could only be del/del, dup/del or dup/dup, which were recovered with the expected mendelian ratio. Because these genotypes could be distinguished using Southern blot analysis, we could unambiguously select homozygous mice for each duplication allele. We prevented the occurrence of 'back *trans*-loxing' in these crosses by using mice that did not carry the *SYCP1/Cre* transgene.

Whole-mount *in situ* hybridizations were carried out on fetuses between 11.5 and 13.5 days old using standard procedures and previously described probes^{25,34}. For skeletal preparations, adult or juvenile animals were processed according to ref. 35.

Received 14 May; accepted 19 September 2002; doi:10.1038/nature01189.

- Rijli, F. M. & Chambon, P. Genetic interactions of Hox genes in limb development: learning from compound mutants. *Curr. Opin. Genet. Dev.* **7**, 481–487 (1997).
- Dolle, P., Izpisua-Belmonte, J. C., Falkenstein, H., Renucci, A. & Duboule, D. Coordinate expression of the murine Hox-5 complex homeobox-containing genes during limb pattern formation. *Nature* **342**, 767–772 (1989).
- Davis, A. P., Witte, D. P., Hsieh-Li, H. M., Potter, S. S. & Capecchi, M. R. Absence of radius and ulna in mice lacking *hoxa-11* and *hoxd-11*. *Nature* **375**, 791–795 (1995).
- Fromental-Ramain, C. *et al.* Specific and redundant functions of the paralogous *Hoxa-9* and *Hoxd-9* genes in forelimb and axial skeleton patterning. *Development* **122**, 461–472 (1996).
- Tabin, C. J. Retinoids, homeoboxes, and growth factors: toward molecular models for limb development. *Cell* **66**, 199–217 (1991).

- Duboule, D. How to make a limb? *Science* **266**, 575–576 (1994).
- Johnson, R. L., Riddle, R. D. & Tabin, C. J. Mechanisms of limb patterning. *Curr. Opin. Genet. Dev.* **4**, 535–542 (1994).
- Dolle, P., Izpisua-Belmonte, J. C., Brown, J. M., Tickle, C. & Duboule, D. Hox-4 genes and the morphogenesis of mammalian genitalia. *Genes Dev.* **5**, 1767–1767 (1991).
- Fromental-Ramain, C. *et al.* *Hoxa-13* and *Hoxd-13* play a crucial role in the patterning of the limb autopod. *Development* **122**, 2997–3011 (1996).
- Zakany, J., Fromental-Ramain, C., Warot, X. & Duboule, D. Regulation of number and size of digits by posterior Hox genes: a dose-dependent mechanism with potential evolutionary implications. *Proc. Natl Acad. Sci. USA* **94**, 13695–13700 (1997).
- Duboule, D. Vertebrate Hox genes and proliferation: an alternative pathway to homeosis? *Curr. Opin. Genet. Dev.* **5**, 525–528 (1995).
- Duboule, D. & Morata, G. Colinearity and functional hierarchy among genes of the homeotic complexes. *Trends Genet.* **10**, 358–364 (1994).
- Herauld, Y., Fraudeau, N., Zakany, J. & Duboule, D. Ulnaless (Ul), a regulatory mutation inducing both loss-of-function and gain-of-function of posterior Hoxd genes. *Development* **124**, 3493–3500 (1997).
- Peichel, C. L., Prabhakaran, B. & Vogt, T. F. The mouse Ulnaless mutation deregulates posterior HoxD gene expression and alters appendicular patterning. *Development* **124**, 3481–3492 (1997).
- Spitz, F. *et al.* Large scale transgenic and cluster deletion analysis of the HoxD complex separate an ancestral regulatory module from evolutionary innovations. *Genes Dev.* **15**, 2209–2214 (2001).
- van der Hoeven, F., Zakany, J. & Duboule, D. Gene transpositions in the HoxD complex reveal a hierarchy of regulatory controls. *Cell* **85**, 1025–1035 (1996).
- Kmita, M., van Der Hoeven, F., Zakany, J., Krumlauf, R. & Duboule, D. Mechanisms of Hox gene colinearity: transposition of the anterior Hoxb1 gene into the posterior HoxD complex. *Genes Dev.* **14**, 198–211 (2000).
- Herauld, Y., Beckers, J., Gerard, M. & Duboule, D. Hox gene expression in limbs: colinearity by opposite regulatory controls. *Dev. Biol.* **208**, 157–165 (1999).
- Herauld, Y., Rassoulzadegan, M., Cuzin, F. & Duboule, D. Engineering chromosomes in mice through targeted meiotic recombination (TAMERE). *Nature Genet.* **20**, 381–384 (1998).
- Duboule, D. Patterning in the vertebrate limb. *Curr. Opin. Genet. Dev.* **1**, 211–216 (1991).
- Dolle, P. *et al.* Disruption of the Hoxd-13 gene induces localized heterochrony leading to mice with neonatal limbs. *Cell* **75**, 431–441 (1993).
- Davis, A. P. & Capecchi, M. R. A mutational analysis of the 5' HoxD genes: dissection of genetic interactions during limb development in the mouse. *Development* **122**, 1175–1185 (1996).
- Kmita, M., Kondo, T. & Duboule, D. Targeted inversion of a polar silencer within the HoxD complex re-allocates domains of enhancer sharing. *Nature Genet.* **26**, 451–454 (2000).
- Nelson, C. E. *et al.* Analysis of Hox gene expression in the chick limb bud. *Development* **122**, 1449–1466 (1996).
- Kondo, T., Zakany, J. & Duboule, D. Control of colinearity in AbdB genes of the mouse HoxD complex. *Mol. Cell* **1**, 289–300 (1998).
- Goff, D. J. & Tabin, C. J. Analysis of Hoxd-13 and Hoxd-11 misexpression in chick limb buds reveals that Hox genes affect both bone condensation and growth. *Development* **124**, 627–636 (1997).
- Foley, K. P., Engel, J. D. *Individual stage selector element mutations lead to reciprocal changes in β - vs. ϵ -globin gene transcription: genetic confirmation of promoter competition during globin gene switching.* *Genes Dev.* **6**, 730–744 (1992).
- Herauld, Y., Beckers, J., Kondo, T., Fraudeau, N. & Duboule, D. Genetic analysis of a Hoxd-12 regulatory element reveals global versus local modes of controls in the HoxD complex. *Development* **125**, 1669–1677 (1998).
- Kmita, M., Tarchini, B., Duboule, D. & Herauld, Y. Evolutionary conserved sequences are required for the insulation of the vertebrate HoxD complex in neural cells. *Development* **129**, 5521–5528 (1976).
- Wijgerde, M., Grosfeld, F. & Fraser, P. Transcription complex stability and chromatin dynamics *in vivo*. *Nature* **377**, 209–213 (1995).
- Dillon, N., Trimborn, T., Strouboulis, J., Fraser, P. & Grosfeld, F. The effect of distance on long-range chromatin interactions. *Mol. Cell* **1**, 131–139 (1997).
- Tanimoto, K., Liu, Q., Bungert, J. & Engel, J. D. Effects of altered gene order or orientation of the locus control region on human beta-globin gene expression in mice. *Nature* **398**, 344–348 (1999).
- Vidal, F., Sage, J., Cuzin, F. & Rassoulzadegan, M. Cre expression in primary spermatocytes: a tool for genetic engineering of the germ line. *Mol. Reprod. Dev.* **51**, 274–280 (1998).
- Herauld, Y., Hrabá-Revev, S., van der Hoeven, F. & Duboule, D. Function of the *Evx-2* gene in the morphogenesis of vertebrate limbs. *EMBO J.* **15**, 6727–6738 (1996).
- Inouye, M. Differential staining of cartilage and bone in fetal mouse skeleton by alcian blue and alizarin red. *S. Cong. Anom.* **16**, 171–173 (1976).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements We thank M. Friedli for technical assistance, and J. Zakany and other colleagues from the Duboule laboratory for sharing mice, reagents, discussions and comments on the manuscript. This work was supported by funds from the Canton de Genève, the Swiss National Research Fund, the Claraz, Latsis, Cloetta and Louis-Jeantet foundations, as well as the NCCR Frontiers in Genetics.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.D. (e-mail: Denis.Duboule@zoo.unige.ch).

Dark cores in sunspot penumbral filaments

Göran B. Scharmer, Boris V. Gudiksen, Dan Kiselman, Mats G. Löfdahl & Luc H. M. Rouppe van der Voort

The Institute for Solar Physics of the Royal Swedish Academy of Sciences, AlbaNova University Centre, SE-106 91 Stockholm, Sweden

Sunspot umbrae—the dark central regions of the spots—are surrounded by brighter filamentary penumbrae, the existence of which remains largely inexplicable¹. The penumbral filaments contain magnetic fields with varying inclinations² and are associated with flowing gas^{3–5}, but discriminating between theoretical models^{6–8} has been difficult because the structure of the filaments has not hitherto been resolved. Here we report observations of penumbral filaments that reveal dark cores inside them. We cannot determine the nature of these dark cores, but their very existence provides a crucial test for any model of penumbrae. Our images also reveal other very small structures, in line with the view that many of the fundamental physical processes in the solar photosphere occur on scales smaller than 100 km.

The best spatial resolution attained in solar observations has typically been about 0.2 arcsec (150 km). The Swedish 1-m Solar Telescope⁹ at the Roque de los Muchachos on the Canary island of La Palma is a newly installed (spring 2002), evacuated refractor designed to significantly improve this. The main obstacle to diffraction-limited imaging is rapidly varying aberrations from temperature inhomogeneities in the Earth's atmosphere. These effects, referred to as 'seeing', are corrected for by low-order adaptive optics (15 corrected modes), real-time frame selection (picking the best images from a continuous stream of exposures), and subsequent image restoration using a variant¹⁰ of the phase-diversity technique^{11,12}. The resolution achieved is better than 0.12 arcsec.

A notable feature in our best images of sunspots is that many penumbral filaments, which are isolated from the bulk of the penumbra and surrounded by dark umbra, show dark cores. The footpoints of the dark cores appear mostly adjacent to—and in some cases, centred on—peripheral umbral dots (small bright features standing out near the edge of the dark umbra) or penumbral grains (brighter parts of penumbral filaments that move radially). These footpoints are in constant motion, sometimes appearing to pass in front of (above), or splitting, the bright dot or grain. Branching of the dark core into several cores can sometimes be seen close to a footpoint (Figs 1 and 2d). The widths of the filaments close to their point of origin are consistently 150–180 km (5–6 image pixels), as measured between the intensity maxima on either side of the dark core on filaments isolated enough to show such maxima. We propose this to be a typical dimension of the structure. The dark cores are unresolved, and could be much narrower than their measured full-width at half-minimum of 90 km. They can be very long compared to their width, and some can be followed for more than 1,000 km (Figs 1, 2 and 3d). The dark-cored filaments can be symmetric (Fig. 2a, d, e), slightly asymmetric (Fig. 2b) or strongly (Fig. 2c) asymmetric.

Deformation of images by 'seeing' can in principle produce spurious double structures. We are, however, confident that the dark cores shown here are real. They look qualitatively the same at the different wavelength bands we have used (430.5 nm, 436.4 nm, 487.7 nm), phase-diversity restored images show them, and they persist over many frames in our best data (see movie in Supplementary Information).

Not all isolated filaments show the dark-core structures but they are found in all spots for which we have excellent data covering at least several minutes. The lifetimes of the dark cores are similar to

the lifetimes of the filaments themselves; they can last for at least one hour.

An interesting variety of dark-cored filaments are the fainter structures in the umbra, of which Fig. 2e shows an example. These features are associated with the inward migration of a bright dot followed by repeated brightening and fading on a timescale of minutes. This suggests that a larger fraction of umbrae than observed so far could have faint or small-scale filamentary structure¹³.

The description above applies to spots close to the centre of the solar disk and thus with a nearly vertical line of sight. We have not found clear and symmetric dark-cored filaments in spots closer to the limb. This could be due to their three-dimensional structure and the oblique viewing angle.

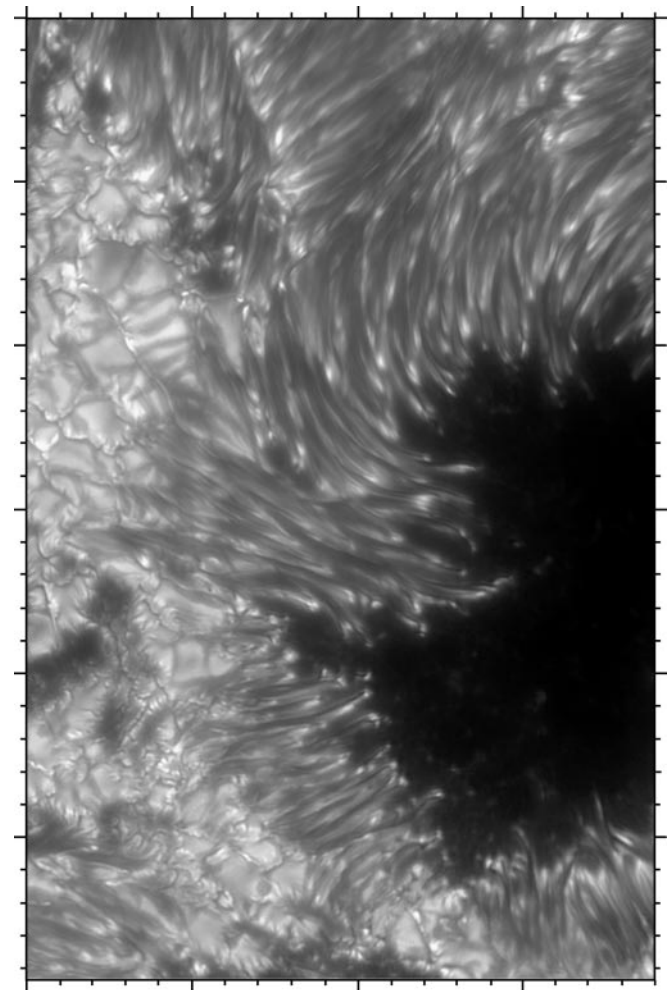


Figure 1 Part of the largest spot in active region 10030 on 15 July 2002 recorded with the Swedish 1-m Solar Telescope using adaptive optics (as are all images shown here). The telescope has an unobstructed aperture of 96 cm and a focal length of 20.3 m at 460 nm wavelength which gives an image scale of 0.041 arcseconds per pixel. The chromatic aberrations from the singlet objective, acting also as vacuum window, are nearly perfectly compensated by an in-vacuum Schupmann corrector. This image was phase-diversity restored, and to that end the camera was equipped with a beamsplitter that puts two images on the same charge-coupled device (CCD) with a known defocus for one of the images. A narrow-band filter centred on 430.5 nm was used. The exposure time was 6 ms. Tickmarks have a spacing of 1,000 km on the Sun. About half of the dark umbra of the sunspot is seen to the right, the surrounding granulation to the left with penumbral structures in between. Many penumbral filaments with dark cores are seen protruding into the umbra, notably in the mass of filaments in the centre of the image. The spectral band used is at the location of the G-band spectral feature (formed by spectral lines of the CH radical), but it should be noted that the dark cores show up also in continuum bands.

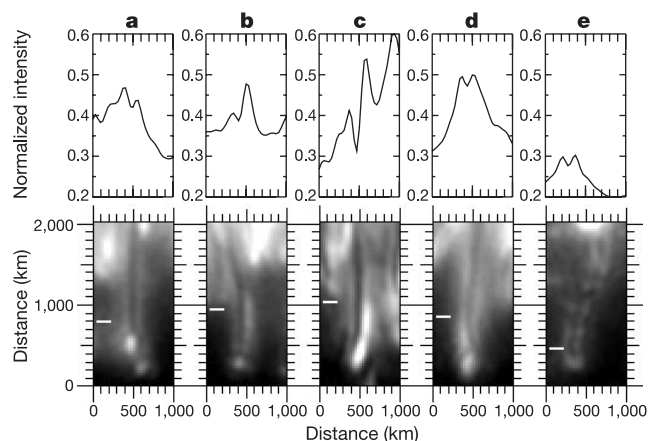


Figure 2 Examples of penumbral filaments with dark cores. Plots (top row) show intensity profiles at indicated sections of the images (bottom row). Intensity values are given in units of the surrounding photosphere from the same exposures. Spatial tickmarks have 100-km spacing. From the largest spot of active region 9957 on 22 May 2002: **a**, G-band (3.3 ms), **b**, G-band (3.3 ms). From the largest spot of active region 10030 on 15 July 2002: **c**, G-band (6 ms). **d**, 487.7-nm continuum (13 ms). **e**, 487.7-nm continuum (13 ms). **a**, **b**, **d** and **e** are corrected for the modulation transfer function (MTF) of the diffraction-limited telescope and the detector by Wiener filtering. Panel **c** is phase-diversity restored, and is thus filtered using a measured optical transfer function.

The data we have obtained so far do not give direct information on fluid velocities or magnetic fields (additional instrumentation is required for this), and thus any interpretation of the dark-cored filaments must be speculative. They could be caused by two (or more) thin bright filaments that happen to pair up in parallel. The consistent separation of these filaments, as well as the very stable behaviour of the dark cores when seen in movies (Supplementary Information), make this explanation unlikely. Such thin (<100 km) and long (>1,000 km) bright filaments would put strong constraints on the mechanisms supplying them with energy without which they would rapidly cool and become invisible¹⁴.

A dark-cored filament could be produced by an optically thin cylindrical tube with hot walls—perhaps a magnetic flux tube heated on the surface by the dissipation of electrical currents. Estimates show, however, that the currents induced by a twist of the tube¹⁵ are insufficient, and topological dissipation¹⁶ (from movements of the flux-tube footpoints) can hardly explain why so few individual flux tubes would be subjected to excess heating for so long.

Inspection of our images shows numerous varieties of other very thin dark lines in magnetic regions that are exemplified in Fig. 3: 'hairs' that are seemingly emanating from pores into the closest neighbouring granules, 'canals' in the granulation near spots and pores, and running dark streaks crossing penumbral filaments diagonally. (These streaks make the filaments look as if they are twisting with minute-long periods and several turns along their length.) Frequently these structures give a visual impression of something dark and opaque lying on top of brighter structures. However, we are cautious about subjective image interpretations. An argument against such top-heavy configurations is that they would be prone to Rayleigh–Taylor instability¹⁷.

Although the sophistication of, and computational resources for, numerical modelling and simulations of complex radiative hydrodynamics and magnetohydrodynamics in the solar atmosphere have improved at an impressive rate during the last decade, it has become increasingly clear that further progress needs observations at higher spatial resolution than is so far achievable. We believe that the observations presented here represent a significant improvement in

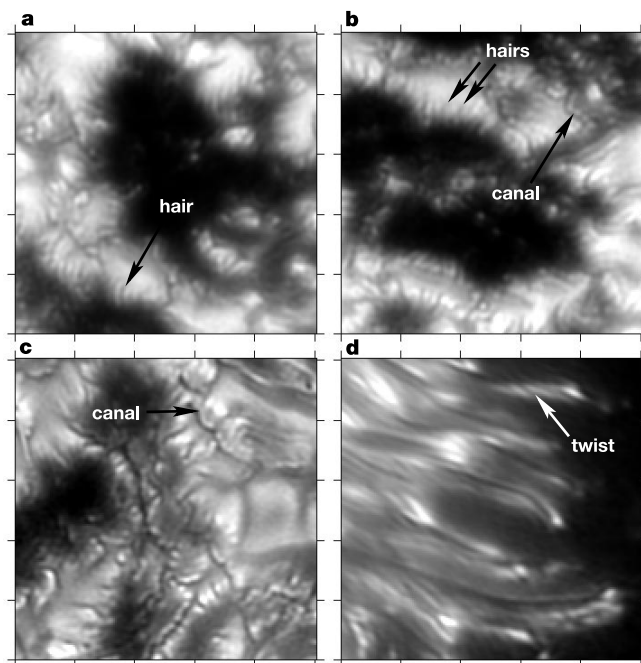


Figure 3 Thin dark features in active region 10030 on 15 July 2002. **a–c**, Pores with hair-like features around their edges and 'canals' in their surroundings. The 'canals' look like more stable and narrow versions of intergranular lanes. **d**, Dark-cored filaments and a filament with dark streaks making it look twisted. Panels **a** and **b** are MTF-filtered 487.7-nm continuum (13 ms); **c** and **d** are phase-diversity restored G-band (6 ms). Tickmarks have 1,000 km spacing.

spatial resolution, while also indicating that fundamental physical scales of solar processes are finally becoming observable. □

Received 19 August; accepted 1 October 2002; doi:10.1038/nature01173.

1. Parker, E. *Sunspots: Theory and Observations* (eds Thomas, J. H. & Weiss, N. O.) 413–423 (Kluwer, Dordrecht, 1992).
2. Title, A. *et al.* On the magnetic and velocity field geometry of a simple sunspot. *Astrophys. J.* **403**, 780–796 (1993).
3. Evershed, J. Radial movement in sun-spots. *Mon. Not. R. Astron. Soc.* **69**, 454–457 (1909).
4. Muller, R. Étude morphologique et cinématique des structures fines d'une tache solaire. *Sol. Phys.* **29**, 55–73 (1973).
5. Sobotka, M., Brandt, P. N. & Simon, G. W. Fine structure in sunspots. III. Penumbral grains. *Astron. Astrophys.* **348**, 621–626 (1999).
6. Danielson, R. E. The structure of sunspot penumbras. II. Theoretical. *Astrophys. J.* **134**, 289–311 (1961).
7. Montesinos, B. & Thomas, J. H. The Evershed effect in sunspots as a siphon flow along a magnetic flux tube. *Nature* **390**, 485–487 (1997).
8. Schlichenmaier, R., Jahn, K. & Schmidt, H. U. Magnetic flux tubes evolving in sunspots. A model for the penumbral fine structure and the Evershed flow. *Astron. Astrophys.* **337**, 897–910 (1998).
9. Scharmer, G. B., Bjelksjö, K., Korhonen, T. K., Lindberg, B. & Pettersson, B. *Innovative Telescopes and Instrumentation for Solar Astrophysics* (eds Keil, S. & Avakyan, S.) Vol. 4853–47 *Proc. SPIE* (in the press).
10. Löfdahl, M. G. *Image Reconstructions from Incomplete Data II* (eds Bones, P. J., Fiddy, M. A. & Milland, R. P.) Vol. 4792–21 *Proc. SPIE* (in the press).
11. Löfdahl, M. G., Berger, T. E. & Seldin, J. H. Two dual-wavelength sequences of high-resolution solar photospheric images captured over several hours and restored by use of phase diversity. *Astron. Astrophys.* **377**, 1128–1135 (2001).
12. Thomas, J. H. Solar physics: The sun at small scales. *Nature* **396**, 114–115 (1998).
13. Livingston, W. Radial filamentary structure in a sunspot umbra. *Nature* **350**, 45–46 (1991).
14. Schlichenmaier, R., Bruls, J. H. M. J. & Schüssler, M. Cooling of a hot flux tube in the solar photosphere. *Astron. Astrophys.* **349**, 961–973 (1999).
15. Galsgaard, K. & Nordlund, Å. Heating and activity of the solar corona. 2. Kink instability in a flux tube. *J. Geophys. Res.* **102**, 219–230 (1997).
16. Parker, E. N. Topological dissipation and the small-scale fields in turbulent gases. *Astrophys. J.* **174**, 499–510 (1972).
17. Spruit, H. C. *The Role of Fine-scale Magnetic Fields on the Structure of the Solar Atmosphere* (eds Schröter, E. H., Vazquez, M. & Wyller, A. A.) 199–209, (Cambridge Univ. Press, Cambridge, 1987).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements The Swedish 1-m Solar Telescope is operated on the island of La Palma by the Royal Swedish Academy of Sciences in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofísica de Canarias.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.K. (e-mail: dan@astro.su.se).

High-power terahertz radiation from relativistic electrons

G. L. Carr*, Michael C. Martin†, Wayne R. McKinney†, K. Jordan‡, George R. Neil‡ & G. P. Williams‡

* National Synchrotron Light Source, Brookhaven National Laboratory, Upton, New York 11973, USA

† Advanced Light Source Division, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA

‡ Free Electron Laser Facility, Jefferson Laboratory, 12000 Jefferson Avenue, Newport News, Virginia 23606, USA

Terahertz (THz) radiation, which lies in the far-infrared region, is at the interface of electronics and photonics. Narrow-band THz radiation can be produced by free-electron lasers¹ and fast diodes^{2,3}. Broadband THz radiation can be produced by thermal sources and, more recently, by table-top laser-driven sources^{4–6} and by short electron bunches in accelerators⁷, but so far only with low power. Here we report calculations and measurements that confirm the production of high-power broadband THz radiation from subpicosecond electron bunches in an accelerator. The average power is nearly 20 watts, several orders of magnitude higher than any existing source, which could enable various new applications. In particular, many materials have distinct absorptive and dispersive properties in this spectral range, so that THz imaging could reveal interesting features. For example, it would be possible to image the distribution of specific proteins or water in tissue, or buried metal layers in semiconductors^{8,9}; the present source would allow full-field, real-time capture of such images. High peak and average power THz sources are also critical in driving new nonlinear phenomena and for pump–probe studies of dynamical properties of materials^{10,11}.

The THz region ($1 \text{ THz} \approx 33 \text{ cm}^{-1}$ or 4 meV) lies in the far-infrared spectral range where conventional thermal sources are very weak. For example, a blackbody source at $2,000 \text{ K}$ provides less than $1 \mu\text{W per cm}^{-1}$ of spectral power density for a typical spectroscopy application. Whereas narrowband sources have been available using free-electron laser (FEL) technology^{1,12}, a significant advance in broadband THz sources has occurred over the past decade with the advent of coherent THz radiation emission from photocarriers in biased semiconductors. Table-top systems using optical rectification of femtosecond lasers either at high repetition rates⁵ or high peak power⁶ are routinely available.

The present work describes a different process for producing coherent THz radiation by accelerated electrons. Like the method described in ref. 5, the process begins with pulsed laser excitation in GaAs, but makes use of photoemission to produce bunches of free electrons in space. Using the energy-recovered linac (ERL) at the Jefferson Laboratory FEL¹³, very short electron bunches ($\sim 500 \text{ fs}$) are brought to relativistic energies ($\sim 40 \text{ MeV}$) in a linac and then transversely accelerated by a magnetic field to produce the desired THz emission as synchrotron radiation. This unique accelerator is capable of a running with a relatively high average beam current (up

to 5 mA). Like the THz emitter described in ref. 5, the electrons experience a common acceleration. If the electron bunch dimensions are small (in particular, the bunch length is less than the wavelength of observation), we again obtain a multiparticle coherent enhancement^{14,15}. Such coherent synchrotron radiation has been observed from electrons accelerated in linacs^{7,16–18}, from compact waveguide FELs¹⁹ and from magnetic undulators^{19–21}. Coherent THz radiation has been discussed, and observed from electron bunches in storage rings^{22–25}, but not yet stably enough for use as a light source. Active programmes to study THz radiation from linacs or storage rings are underway at BESSY II (Berliner Elektronenspeicherring-Gesellschaft für Synchrotronstrahlung; www.bessy.de) and DESY (Deutsches Elektronen Synchrotron) in Germany, and at Brookhaven National Laboratory and Lawrence Berkeley National Laboratory in the USA. In addition, programmes are underway at ENEA-Frascati to generate broadband THz radiation by exploiting the distinctive properties of waveguide FELs which arise when the electron velocity is close to the group velocity of the wave packet²⁶. Some linacs can create very short bunches ($< 1 \text{ ps}$) and produce coherent radiation up to a few THz, but most are limited to repetition rates of a few Hz, so the average power is quite low. The repetition rate for storage rings is of the order of 100 MHz , but the electron bunches are significantly longer ($\sim 100 \text{ ps}$) owing to longitudinal damping through synchrotron radiation emission. Thus the emission is limited to the very low frequency regime (far-infrared), or arises from instabilities that briefly modify the bunch shape.

Our ERL accelerator system overcomes some of the limitations of conventional linacs and storage rings. Electron bunches as short as $\sim 500 \text{ fs}$ are produced by the standard technique of energy modulation (chirping) followed by compression in the dispersive region of a magnetic chicane²⁷. The time taken for an electron bunch to pass through the accelerator is less than $1 \mu\text{s}$; thus longitudinal damping is negligible. But unlike most linacs, our system operates at a very high repetition rate (up to 75 MHz) by using superconducting radio frequency cavities and recovering the energy of the spent electron bunches¹³ so that the average current is orders of magnitude higher than in conventional linacs.

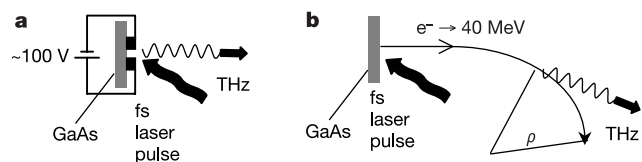


Figure 1 Comparison between coherent THz radiation generated by an 80-MHz conventional laser-driven THz source (a) and the relativistic source described here (b). In a, the photo-induced carriers immediately experience a force from the bias field ($\sim 100 \text{ V}$ across a $100 \mu\text{m}$ gap) of $\sim 10^6 \text{ V m}^{-1}$, which results in an acceleration of 10^{17} m s^{-2} . The entire process is completed in less than 1 ps , resulting in spectral content up to a few THz. In b, approximately the same number of charge carriers are brought to a relativistic energy of $> 10 \text{ MeV}$ in a linac, after which a magnetic field bends their path into a circle of radius $\rho = 1 \text{ m}$, resulting in an acceleration $c^2/\rho = 10^{17} \text{ m s}^{-2}$, the same as for a. An observer of b would also detect a brief pulse of electromagnetic radiation as an electron bunch passed by. But in this case, two factors control the pulse duration; one is the bunch length, and the other is the time for the relativistically compressed acceleration field from each electron to sweep past. The latter is given approximately by $\delta t = 4\rho/(3\gamma^3c)$, and determines the spectral range emitted by each electron. The bunch length determines the spectral range over which the coherent enhancement occurs. For an electron energy of 10 MeV ($\gamma = 21$), and with $\rho = 1 \text{ m}$, we obtain a δt of about 500 fs , which is comparable to the bunch length. The resulting spectral content extends up to about 1 THz , the same spectral range as for a. With all factors except γ the same, we see from equation (1) that the power radiated by a relativistic electron exceeds that from a conventional THz emitter by a factor of $\gamma^4 = 21^4 = 2 \times 10^5$.

We now examine how our system generates THz power many orders of magnitude more efficiently than a more conventional (non-relativistic) THz source⁵. The diagram in Fig. 1 shows the two processes for comparison. In both cases, a short light pulse from a mode-locked laser strikes a GaAs wafer, generating charge carriers. Thus the number of radiating charges should be comparable (same order of magnitude) for both cases. We can therefore compare the power produced per electron, and use Larmor's formula²⁸ for the radiated power P . In c.g.s. units it takes the form

$$P = \frac{2e^2 a^2}{3c^3} \gamma^4 \quad (1)$$

where e is the charge, a is the acceleration, c the speed of light and γ is the ratio of the mass of the electron to its rest mass. With all factors except γ the same, we see from equation (1) that for our case the power radiated by a relativistic electron exceeds that from a conventional THz emitter by a factor of $\gamma^4 = 21^4 = 2 \times 10^5$. In practice, the electron energy can be significantly larger, but this simply adds intensity at higher frequencies and leaves the low-frequency (THz) intensity essentially unchanged.

We mention again that both cases benefit from multiparticle coherent emission, as many radiating charges are contained physically within one period of the emitted THz radiation. Theoretically, the more general expression for the power emitted by an electron bunch, as a function of frequency (ω) and solid angle (Ω), is derived by extending the classical theory of electrodynamics²⁸ for a single electron, to a system of N electrons, thus^{14,15}:

$$\frac{d^2 I}{d\omega d\Omega} = [N[1 - f(\omega)] + N^2 f(\omega)] \times \left| \frac{e^2 \omega^2}{4\pi^2 c} \int_{-\infty}^{\infty} \hat{n} \times (\beta \times \hat{n}) e^{i\omega(t - \frac{\mathbf{r} \cdot \hat{n}}{c})} dt \right|^2 \quad (2)$$

where β is the ratio of the velocity of the particle bunch to the velocity of light, $\hat{n}$ is a unit vector along the direction of propagation (to the observer), $\mathbf{r}(t)$ is the location of the electron bunch centre, and N is the number of particles in the bunch. The term $N^2 f(\omega)$ represents the coherent enhancement, and includes the form factor $f(\omega)$ which is the Fourier transform of the normalized longitudinal particle distribution within the bunch; that is,

$$f(\omega) = \left| \int_{-\infty}^{\infty} e^{i\omega \hat{n} \cdot \mathbf{z}/c} S(z) dz \right|^2 \quad (3)$$

where $S(z)$ is the distribution function for particles in the bunch, measured relative to the bunch centre.

A practical solution to the second major term in equation (2) has been presented in ref. 29. For the present calculations (Fig. 2), we assume 40-MeV electron bunches each carrying 100 pC of charge, passing through a 1-m-radius bend at a 37.4-MHz repetition rate. For simplicity, we assume that each bunch has a gaussian particle distribution of width σ , yielding a gaussian form factor:

$$f(\omega) = e^{-\left(\frac{\omega \sigma}{c}\right)^2} = e^{-4\pi^2 \sigma^2 \left(\frac{\lambda}{\lambda_0}\right)^2} \quad (4)$$

where λ is the wavelength of the light at frequency ω . The bunches are not strictly gaussian in practice, but this approximation is useful for estimating the overall power and spectral content. The electron beam had an r.m.s. width and height of 1 mm. Given a natural emission angle, $1.66(\lambda/\rho)^{1/3}$ of 0.11 rad at 1 THz, where ρ is the bending radius, the diffraction-defined source size was of the order of the electron beam size, and the emitted radiation had an extremely high degree of transverse spatial coherence. This allows for interferometry by a simple form of wavefront division³⁰.

In our experiments using the ERL THz source, the electrons were generated using the frequency-doubled output of a Nd:YLF laser (model Antares, made by Coherent) operating at, or a sub-multiple of, 74.8 MHz, and with an average power of a few watts. Light of

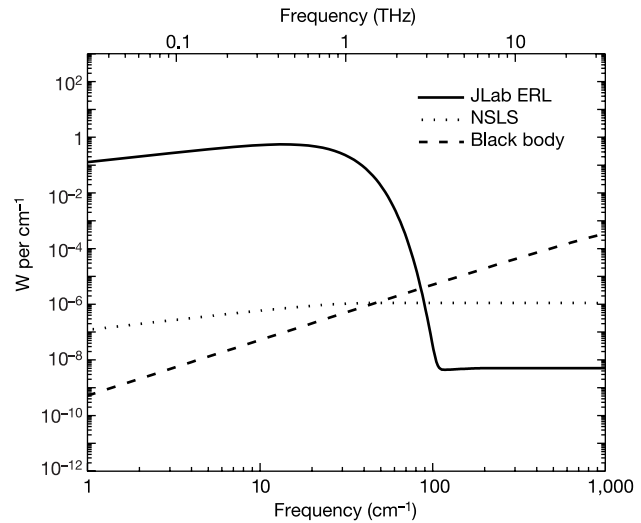


Figure 2 Calculations of the average power emitted by a 10-mm² thermal source at 2,000 K (dashed line), the NSLS VUV ring at Brookhaven National Laboratory (dotted line), and the Jefferson Laboratory (JLab) ERL (solid line). The NSLS was calculated for a stored current of 800 mA and 90 mrad (vertical) by 90 mrad (horizontal) collection. The JLab ERL was calculated for 500-fs FWHM, 100-pC bunches at a repetition rate of 37.4 MHz (for an average current of 3.7 mA) and 60 mrad (vertical) by 60 mrad (horizontal) collection. The increased power of the JLab ERL in the THz range is clear. We also note that the calculation assumes that the ERL has an average current of 3.7 mA compared with 800 mA at the NSLS, Brookhaven, so that in the incoherent regime the NSLS has 800/3.7 = 216 times the power per cm⁻¹. However, in the THz spectral range (where the JLab ERL emits coherently and the NSLS does not) the ERL has 3.9×10^6 times the power of the NSLS.

wavelength 530 nm was incident on a negative electron affinity Cs-coated GaAs cathode. The resulting photoelectrons were accelerated using a d.c. voltage of 300 kV into a superconducting linac, and accelerated to an energy of 40 MeV. Although the electrons are initially emitted from the cathode with a pulse length of ~ 40 ps full-width at half-maximum, they become tightly bunched in the accelerator to pulse lengths less than 1 ps. After passing through the accelerator system, the electrons are decelerated in the same linac to an energy of 10 MeV before reaching the beam dump, thus recovering most of the beam energy. This energy recovery allows average current of up to 5 mA and electron bunches containing up to 135 pC, using an r.f. system nominally capable of accelerating only 1.1 mA beam current.

The ERL THz radiation was extracted from a dipole magnet of 1 m bending radius immediately before the FEL cavity, the latter being unimportant for these experiments. For the total power measurements, the radiation left the accelerator vacuum chamber through a 10-mm-aperture diamond window subtending an angle of 20×20 mrad relative to the source point. The emerging beam was focused onto a calibrated LiTaO₃ pyroelectric detector, calibrated with equipment traceable to NIST. This detector had a nearly flat response (J25, Molelectron) out to THz wavelengths owing to a black organic coating, and a nominal responsivity of 8.83 V J^{-1} ($\pm 2\%$).

The spectral content of the ERL THz radiation was analysed using a rapid-scan Michelson interferometer (Nexus 670, Nicolet) with a silicon beamsplitter. The light was detected using a 4.2 K bolometer (Infrared Laboratories) with a 2 mm $\times$ 2 mm boron-doped Si composite element, fed from a 12-mm-diameter $f/4$ Winston cone. It was fitted with a black polyethylene filter to ensure no radiation above 600 cm^{-1} was detected. The diamond window on the accelerator was replaced by a larger crystal-quartz window to

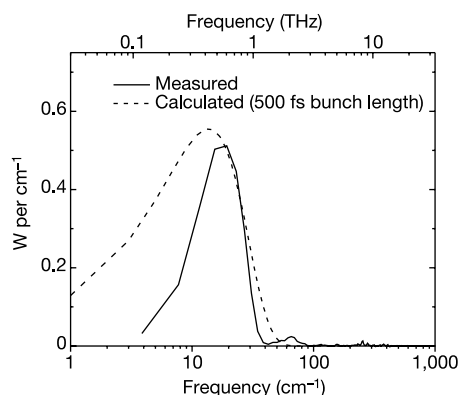


Figure 3 Comparison between measured (solid line) and calculated (dashed line) THz spectral intensity. We also show a curve calculated for a bunch length of 500 fs, which is reasonable for the machine operating under the conditions of the experiment. We have scaled the data to fit the theory on the basis of the absolute power measurements. The spectral onset of the super-radiant enhancement on the high-frequency side is clearly seen, with the shape matching the theoretical prediction. Note that there is a severe discrepancy on the lower-frequency side owing to diffraction effects. This can be understood in the following way. At 10 cm^{-1} and with an $f/17$ beam, the diffraction-limited source size is 17 mm, almost the same as the extraction optics. At 1 cm^{-1} , the diffraction-limited source size would, at 170 mm, be more than 3 times larger than the vacuum chamber containing the electron beam. Note also the additional peak at 60 cm^{-1} , which we attribute to the non-gaussian nature of the electron bunch. As we measured the intensity, not field, of the emitted light, we were unable to determine uniquely the electron density distributions, but this might be possible in future experiments using coherent detection.

increase the energy collection to $60 \times 60\text{ mrad}$. A spherical mirror of 80 cm focal length produced a 48-mm-diameter collimated beam compatible with the interferometer optics. A switching mirror allowed a remote choice of source, namely the THz energy from the accelerator, or a $T = 1,300\text{ K}$ thermal reference source (see below).

For the spectroscopy experiments, the analysis and detection system did not have sufficient dynamic range to cover the seven decades in power difference between the two sources. But as mentioned earlier, the ERL THz source could be run at a precisely defined lower repetition rate. In this way we could reduce the average power without changing the spectral content. We chose to make measurements at 584 kHz, instead of 37.4 MHz, and at a charge per bunch of 34 pC instead of 100 pC, thereby reducing the ERL THz power by a factor of $\{(34 \times 10^6)/(584 \times 10^3)\} \times (100/34)^2$, or approximately 550.

We have another reference point for determining the absolute power, as we were able to switch sources from the ERL THz emission port to a 1,300 K thermal source (the spectrometer's standard 'global' source). This allowed us to measure the relative power using the same spectrometer and detection system. At a frequency of 12 cm^{-1} we obtained a ratio of intensity from the ERL THz source to that of the global of 2×10^4 . To compare with the calculation, we multiply the results for the THz source by the reduction factor of 550, as discussed earlier. This implies a measured advantage of the ERL THz source over the global of 10^7 . The calculation predicts an enhancement of $0.6/(6 \times 10^{-8}) = 10^7$. The data are shown in Fig. 3, and the result affirms the large ERL THz power. The level of agreement is somewhat surprising, as our simple arguments have ignored diffraction and detection efficiency. In fact, owing to uncertainties in the absorbance of the detector coating in the THz region, the reported data place a lower bound on the power measurement.

One additional property of super-radiant emission from elec-

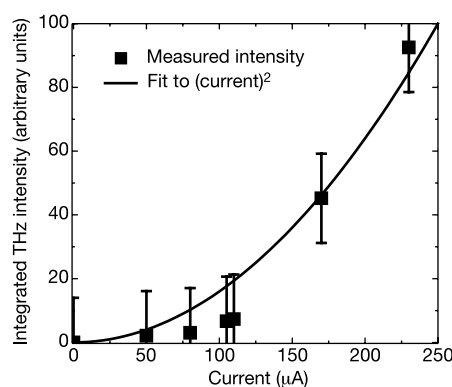


Figure 4 Measured THz intensity as a function of beam current (square symbols), showing the quadratic dependence expected for coherent emission (solid line). We note that in our experiments we were able to observe considerable changes in spectral weight depending on operating conditions, and we were able to enhance certain spectral regions in a controlled way via the machine parameters. Thus the electron bunch distributions were not purely gaussian. This accounts for the large error bars.

trons is the dependence of the intensity on the square of the number of particles per bunch from equation (2). In Fig. 4 we plot the integrated intensity as a function of bunch charge, which shows good agreement with the theoretical N^2 curve.

Finally, we measured the polarization of the emitted THz radiation. The intensity ratio for the horizontal to vertical polarization components is 3 for synchrotron radiation in the long-wavelength limit. This assumes full collection of the emitted radiation. We note that the dominant intensity is near 30 cm^{-1} , which has a natural opening angle of 86 mrad. As the emission pattern is 'clipped' by the 60-mrad collection optics, the calculated ratio is expected to be higher, approaching a value of 6. Using a wire-grid polarizer placed between the Michelson modulator and the detector, we measured a ratio of 5 and consider this to be good agreement.

We have produced broadband, high-brightness, THz radiation with close to 1 W per cm^{-1} of average spectral power density into the diffraction limit, and peak spectral power densities about 10^4 times higher than this. However, one area of concern remains—the question of timing jitter and current fluctuations, and their effect on coherent detection and signal to noise ratio. As energy-recovery linac THz sources are being considered for other applications, some of which involve timing, it is important to evaluate timing jitter and current fluctuations. We have made preliminary measurements using a frequency analyser, and plan further detailed studies. □

Received 12 March; accepted 30 September 2002; doi:10.1038/nature01175.

- Ramian, G. The new UCSB free-electron lasers. *Nucl. Instrum. Methods Phys. Res. A* **318**, 225–229 (1992).
- Porterfield, D. W., Crowe, T. W., Bradley, R. F. & Erickson, N. R. A high-power, fixed-tuned, millimeter-wave balanced frequency doubler. *IEEE Trans. Microwave Theory Techn.* **MTT-47**, 419–425 (1999).
- Siegel, P. H. Terahertz technology. *IEEE Trans. Microwave Theory Techn.* **MTT-50**, 910–928 (2002).
- Auston, D. H., Cheung, K. P., Valdmann, J. A. & Kleinman, D. A. Cherenkov radiation from femtosecond optical pulses in electrooptic media. *Phys. Rev. Lett.* **53**, 1555–1558 (1984).
- Bonvalet, A., Joffe, M., Martin, J. L. & Migus, A. Generation of ultrabroadband femtosecond pulses in the mid-infrared by optical rectification of 15 fs light pulses at 100 MHz repetition rate. *Appl. Phys. Lett.* **67**, 2907–2909 (1995).
- You, D., Jones, R. R., Bucksbaum, P. H. & Dykaar, D. R. Generation of high-power sub-single-cycle 500-fs electromagnetic pulses. *Opt. Lett.* **18**, 290–292 (1993).
- Nakazato, T. *et al.* Observation of coherent synchrotron radiation. *Phys. Rev. Lett.* **63**, 1245–1248 (1989).
- Chen, Q. & Zhang, X.-C. *Ultrafast Laser Technology and Applications* (eds Fermann, M. E., Galvanauskas, A. & Sucha, G.) 521–572 (Marcel Dekker, New York, 2001).
- Zhang, X.-C. in *Compound Optoelectronic Materials and Devices* 69–80 (1995).
- Cole, B. E., Williams, J. B. & King, B. T. Coherent manipulation of semiconductor quantum bits with terahertz radiation. *Nature* **410**, 60–63 (2001).
- Huber, R. *et al.* How many-particle interactions develop after ultrafast excitation of an electron-hole plasma. *Nature* **414**, 286–289 (2001).

12. Winnerl, S. *et al.* Frequency doubling and tripling of terahertz radiation in a GaAs/AlAs superlattice due to frequency modulation of Bloch oscillations. *Appl. Phys. Lett.* **77**, 1259–1261 (2000).
13. Neil, G. R. *et al.* Sustained kilowatt lasing in a free-electron laser with same cell energy recovery. *Phys. Rev. Lett.* **84**, 662–665 (2000).
14. Nodvick, S. & Saxon, D. S. Suppression of coherent radiation by electrons in a synchrotron. *Phys. Rev.* **96**, 180–184 (1954).
15. Hirschmugl, C. J., Sagurton, M. & Williams, G. P. Multiparticle coherence calculations for synchrotron radiation emission. *Phys. Rev. A* **44**, 1316–1320 (1991).
16. Happek, U., Blum, E. B. & Sievers, A. J. Observation of coherent transition radiation. *Phys. Rev. Lett.* **67**, 2962–2965 (1991).
17. Wang, D. X., Krafft, G. A. & Sinclair, C. K. Measurement of femtosecond electron bunches using a rf zero-phasing method. *Phys. Rev. E* **57**, 2283–2286 (1998).
18. Lihn, H.-C., Kung, P., Settakorn, C. & Wiedemann, H. Observation of stimulated transition radiation. *Phys. Rev. Lett.* **76**, 4163–4166 (1996).
19. Doria, A., Bartolini, R., Feinstein, J., Gallerano, G. P. & Pantell, R. H. Coherent emission and gain from a bunched electron beam. *IEEE J. Quant. Electron.* **QE-29**, 1428–1436 (1993).
20. Jaroszynsky, D. A., Bakker, R. J., van der Geer, C. A. J., Oepts, D. & van Amersfoort, P. W. Coherent start-up of an infrared free electron laser. *Phys. Rev. Lett.* **71**, 3798–3801 (1993).
21. Berryman, K. W., Crosson, E. R., Ricci, K. N. & Smith, T. I. Coherent spontaneous radiation from highly bunched electron beams. *Nucl. Instrum. Methods Phys. Res. A* **375**, 526–529 (1996).
22. Tamada, H., Tsutsui, H., Shimoda, K. & Mima, K. Features of the compact photon storage ring. *Nucl. Instrum. Methods A* **331**, 566–571 (1993).
23. Andersson, A., Johnson, M. S. & Neland, B. Coherent synchrotron radiation in the far infrared from a 1-mm electron bunch. *Opt. Eng.* **39**, 3099–3105 (2000).
24. Arp, U. *et al.* Spontaneous coherent microwave emission and the sawtooth instability in a compact storage ring. *Phys. Rev. Spec. Topics Accelerat. Beams* **4**, 54401 (2001).
25. Carr, G. L. *et al.* Observation of coherent synchrotron radiation from the NSLS VUV ring. *Nucl. Instrum. Methods A* **463**, 387–392 (2001).
26. Giovenale, E. *et al.* Longitudinal phase-space manipulation: a device to enhance the coherent emission from an RF modulated electron beam. *Nucl. Instrum. Methods* **437**, 128–133 (1999).
27. Yu, L. H., Johnson, E., Li, D. & Umstadter, D. Femtosecond free-electron laser by chirped pulse amplification. *Phys. Rev. E* **49**, 4480–4486 (1994).
28. Jackson, J. D. *Classical Electrodynamics* (Wiley, New York, 1975).
29. Hulbert, S. L. & Williams, G. P. *Handbook of Optics: Classical, Vision, and X-Ray Optics* Vol. III, *Synchrotron Radiation Sources*, 2nd edn (eds Bass, M., Enoch, J. M., Van Stryland, E. W. & Wolfe, W. L.) 32.1–32.20 (McGraw-Hill, New York, 2001).
30. Moeller, K. D. *et al.* *Appl. Opt.* **30**, 4297–4301 (1991).

Acknowledgements We thank our colleagues for their help and support, which were essential for these experiments. This work was supported primarily by the US Department of Energy. The Jefferson Laboratory FEL is supported by the Office of Naval Research, the Air Force Research Laboratory, the Commonwealth of Virginia and the Laser Processing Consortium.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.P.W. (e-mail: gwyn@mailaps.org).

Ferromagnetism of a graphite nodule from the Canyon Diablo meteorite

J. M. D. Coey*, M. Venkatesan*, C. B. Fitzgerald*, A. P. Douvalis* & I. S. Sanders†

* Physics Department, † Geology Department, Trinity College, Dublin 2, Ireland

There are recent reports of weak ferromagnetism in graphite^{1,2} and synthetic carbon materials³ such as rhombohedral C₆₀ (ref. 4), as well as a theoretical prediction of a ferromagnetic instability in graphene sheets⁵. With very small ferromagnetic signals, it is difficult to be certain that the origin is intrinsic, rather than due to minute concentrations of iron-rich impurities. Here we take a different experimental approach to study ferromagnetism in graphitic materials, by making use of meteoritic graphite, which is strongly ferromagnetic at room temperature. We examined ten samples of extraterrestrial graphite from a nodule in the Canyon Diablo meteorite. Graphite is the major phase in every sample, but there are minor amounts of magnetite, kamacite, akaganéite, and other phases. By analysing the phase composition of a series of samples, we find that these iron-rich minerals can only account for about two-thirds of the observed

magnetization. The remainder is somehow associated with graphite, corresponding to an average magnetization of 0.05 Bohr magnetons per carbon atom. The magnetic ordering temperature is near 570 K. We suggest that the ferromagnetism is a magnetic proximity effect induced at the interface with magnetite or kamacite inclusions.

Many carbon-based ferromagnets order magnetically below 20 K (refs 6, 7), but the recent study⁴ of polymerized rhombohedral C₆₀ found a Curie temperature, T_C , of about 500 K and a spontaneous magnetization $\sigma_s = 0.09 \text{ A m}^2 \text{ kg}^{-1}$. An extensive study of graphite samples of different provenance¹ suggested an intrinsic origin for the ferromagnetism, although the spontaneous magnetization at room temperature did not exceed $0.003 \text{ A m}^2 \text{ kg}^{-1}$. Another intriguing report is that ferromagnetism with $\sigma_s = 0.02 \text{ A m}^2 \text{ kg}^{-1}$ may coexist with superconductivity in the graphite–sulphur system⁸. To put the weakness of the ferromagnetism in perspective, carbon with a ferromagnetic moment of 1 Bohr magneton (μ_B) per atom would have $\sigma_s = 465 \text{ A m}^2 \text{ kg}^{-1}$, and a corresponding polarization $J_s = 1.2 \text{ T}$. It seems that either a tiny fraction of the carbon atoms participate in the magnetism of these materials, or the carbon moment is remarkably weak ($\sim 10^{-4} \mu_B$). An exception is amorphous carbon prepared by direct pyrolysis³, which in one case⁹ was reported to have a room-temperature magnetization of $9.2 \text{ A m}^2 \text{ kg}^{-1}$, or $0.02 \mu_B$ per carbon. We have reproduced this result, but find iron in the form of micrometre-sized oxide particles dispersed throughout the carbon.

The impact some 50,000 years ago of the 50,000-tonne Canyon Diablo meteorite at a relative velocity of about 20 km s^{-1} was a cataclysmic event, which created a crater of diameter 1.3 km in the Arizona desert. Canyon Diablo is classified as a silicate-bearing IAB iron^{10,11}. The origin of this group is puzzling, but it may have involved catastrophic mixing of the molten iron core of an asteroid in a collision with a chondritic body in the first few million years of

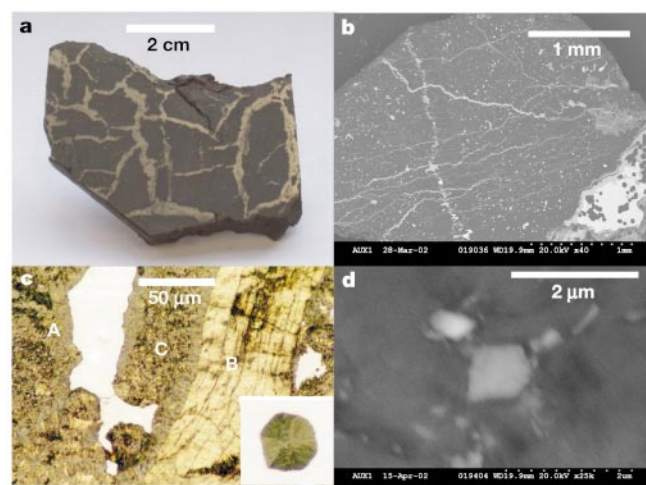


Figure 1 The Canyon Diablo graphite nodule at increasing magnifications. **a**, Cut and polished section of the whole nodule, showing interconnecting veins of kamacite ($\text{Fe}_{94}\text{Ni}_6$) in the graphite matrix. **b**, Backscattered electron image of graphite lightly peppered with tiny nuggets of kamacite (bright flecks) and traversed by thin, subparallel veins of oxidized iron (pale grey). A partially oxidized kamacite vein (bottom right) contains black inclusions of cliftonite¹⁸, a variety of graphite comprising radiating clusters of crystallites that appear to have grown by exsolution from carbon-rich metal on cooling. **c**, Reflected polarized light image of polished graphite showing three distinct forms: A, a continuous border of cliftonite surrounding an embayed nugget of kamacite, B, a large buckled plate of graphite (pale gold), and C, a poorly polished, felted mass of graphite crystallites. Inset, a cliftonite inclusion in metal at the same scale. **d**, Backscattered image of graphite with embedded iron-rich inclusions, possibly magnetite, down to 50 nm in size.

Table 1 Ferromagnetic properties and related data for samples from the Canyon Diablo graphite nodule

Sample	ρ (kg m ⁻³)	σ_s (A m ² kg ⁻¹)	LOI (%)	C (wt%)	Fe (wt%)	Fe ₃ O ₄ (wt%)	Fe ₉₄ Ni ₆ (wt%)	(Fe _{2.0} Ni _{1.0})P + 'nc iron' (wt%)	σ_{imp} (A m ² kg ⁻¹)	$\Delta\sigma$ (A m ² kg ⁻¹)	σ_c (A m ² kg ⁻¹)
1.1	ND	70.0	51.8	54.2	31.8	26.2	6.3	0.0	33.3	36.7	67.7
1.2	ND	30.0	43.9	48.5	36.4	27.0	0.0	9.8	30.1	-	-0.2
1.3	2,630	30.5	74.9	69.8	18.8	6.1	7.8	0.4	21.8	8.7	12.5
1.4	3,040	28.7	46.5	51.5	32.0	15.6	3.6	9.1	28.6	0.1	0.2
1.5	3,010	41.5	46.2	46.2	43.5	15.8	7.2	9.1	36.6	4.9	10.6
1.6	2,430	25.9	62.6	66.1	23.3	15.4	4.3	3.6	24.5	1.4	2.1
1.7	3,350	51.5	28.5	39.8	44.0	16.5	6.2	6.4	32.2	19.3	48.5
1.8	3,380	38.4	51.8	54.4	29.0	9.6	5.2	3.6	22.0	16.4	30.1
1.9	3,360	62.3	29.0	36.7	44.8	19.6	9.8	8.3	44.2	18.1	49.3
1.10	3,650	21.2	53.8	56.0	28.5	11.0	3.2	4.6	19.8	1.4	2.5
Average	3,106	40.6		52.1	34.2				29.7	10.8	23.1
(s.d.)	(412)	(13.1)		(11.7)	(10.5)				(7.0)	(9.5)	(21.5)

The averages are weighted by sample mass. 'nc iron', nanocrystalline iron; ρ , density; σ_s , spontaneous magnetization; LOI, loss on ignition; σ_{imp} , total magnetization of ferromagnetic inclusions; $\Delta\sigma = \sigma_s - \sigma_{imp}$; σ_c , magnetization attributed to carbon. ND, not determined.

the Solar System. The IAB group exhibit great heterogeneity in their range of inclusions. Rounded, graphite-rich nodules are a particular feature of Canyon Diablo. The specimens that we have studied were taken from the nodule illustrated in Fig. 1. Graphite is the major phase, but the nodule contains primary kamacite (Fe₉₄Ni₆) and magnetite (Fe₃O₄), together with traces of schreibersite (Fe_{2.0}Ni_{1.0})P, troilite (FeS) and enstatite (MgSiO₃). The metal has been locally oxidized to magnetite and akaganéite (Cl-containing β -FeO(OH)), with traces of haematite (Fe₂O₃) in some specimens. These phases replace the metal, and also fill stress fractures induced by expansion during terrestrial weathering.

Ten graphite samples of 100–500 mg were investigated. Overall chemical analysis gave (in wt%) C 51.8; S 0.09; P 0.12; FeO_{1.12} 41.6 (the average iron oxidation state comes from Mössbauer data); NiO 2.50; CoO 0.21; SiO₂ 1.70; Al₂O₃ 0.19; MnO < 0.01; MgO 0.49; CaO 0.06; Na₂O 0.02; K₂O 0.02; and TiO₂ 0.01. Densities of 2,400–3,600 kg m⁻³ are all greater than the X-ray density of graphite (2,260 kg m⁻³), which is consistent with the presence of the phases seen in Fig. 1. Part of each sample was ground to powder in an agate mortar for chemical and magnetic analysis. Any obvious kamacite nuggets were removed when grinding the powder. Small fragments

(1–10 mg) were selected from several of the samples in order to investigate the variability of the magnetization.

Each part of every sample is strongly magnetic, jumping to a small ferrite magnet. Magnetization curves are all similar to the one illustrated in Fig. 2. Values of σ_s for the ten powder samples range from 21 to 70 A m² kg⁻¹. The range of magnetization of the milligram fragments was 6–185 A m² kg⁻¹, the latter being measured on a kamacite nugget. The magnetization data are summarized in Table 1. Curie temperatures were determined from thermogravimetric scans in a magnetic field gradient (Fig. 2). The two sharp transitions at 1,030 K and 860 K correspond to T_C for kamacite and magnetite, respectively, and the broad decrease in magnetization at around 570 K is attributed to ferromagnetic graphite.

In order to estimate the magnetization associated with ferromagnetic impurity phases dispersed throughout the graphite, we characterized these phases using Mössbauer spectroscopy, together with chemical analysis and scanning electron microscopy with energy-dispersive X-ray analysis (EDAX). X-ray diffraction (Fig. 3) provides a qualitative overview. Of the iron-rich minerals, magnetite, kamacite and schreibersite are intrinsically ferromagnetic or ferrimagnetic, troilite and haematite are antiferromagnetic, and akaganéite is paramagnetic at room temperature. There is no doubt that the first three minerals contribute to the magnetization of meteoritic graphite, but the question is whether they are present in sufficient quantity to explain the magnetization data in Table 1.

Typical Mössbauer spectra are shown in Fig. 4. Besides the magnetically split components due to magnetite and kamacite

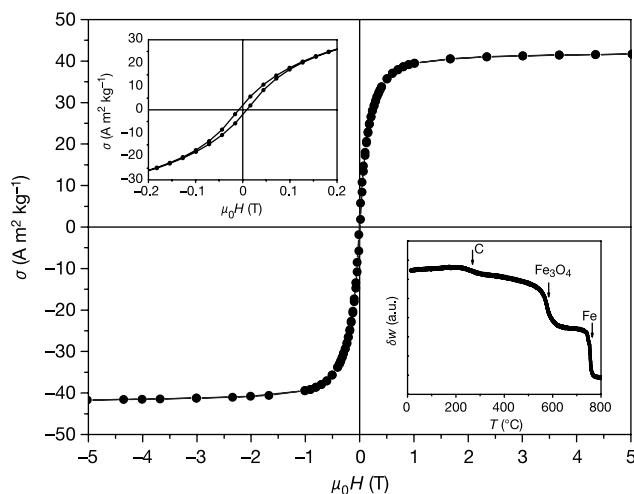


Figure 2 Typical room-temperature curve of magnetization σ against applied field $\mu_0 H$ for graphitic material from the Canyon Diablo meteorite. Insets display the hysteresis (top left) and a thermomagnetic scan showing the Curie temperatures of graphite, magnetite and kamacite (bottom right). There is a little hysteresis, with coercivity $\mu_0 H_c$ in the range 6–8 mT. The low-field tangent to the magnetization curves intersects the line $\sigma = \sigma_s$ at a field $\mu_0 H \approx 200$ mT, which is consistent with the presence of spherical particles with a demagnetizing factor $N \approx 1/3$ and $J_s \approx 0.6$ T.

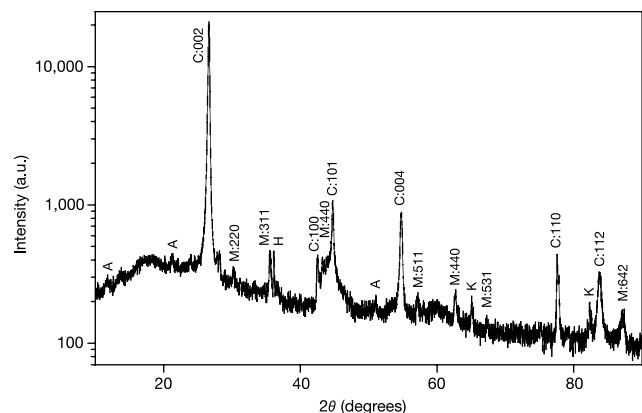


Figure 3 Typical X-ray diffraction pattern for graphitic material from the Canyon Diablo meteorite. The peaks are assigned to graphite/cliftonite (C), magnetite (M), kamacite (K), akaganéite (A) and haematite (H).

and the paramagnetic akaganéite doublet, there is also a broad, poorly resolved component with hyperfine field $B_{\text{hf}} \approx 20$ T, which could be due to poorly crystallized ferric hydroxide or possibly cohenite (Fe_3C). However, low-temperature spectra are fully split, with no component with $B_{\text{hf}} > 30$ T (Fig. 4b). We assume that the recoilless fractions of all phases are similar, so that the relative

amounts of iron in the phases are proportional to their Mössbauer absorption areas.

In Table 1, the concentration (in wt%) of the ferromagnetic minerals in each of the graphitic samples is listed, together with their combined contribution to the magnetization, σ_{imp} , calculated by assuming values of 75, 216 and $100 \text{ A m}^2 \text{ kg}^{-1}$ for oxidized magnetite, kamacite and schreibersite, respectively. The concentration of these minerals was derived by distributing the total iron among the phases in proportion to their Mössbauer absorption areas (or, in the case of $(\text{Fe}_{2.0}\text{Ni}_{1.0})\text{P}$, on the basis of the P content). Experimental uncertainties are given in parentheses. Assuming that the broad component is entirely due to a nanocrystalline iron phase with magnetization of $100 \text{ A m}^2 \text{ kg}^{-1}$, the observed magnetization, σ_s , in Table 1 significantly exceeds σ_{imp} in six out of ten samples (Fig. 4). The average value of σ_s weighted by sample mass ($40.6 \text{ A m}^2 \text{ kg}^{-1}$) is 38% greater than that of σ_{imp} ($29.7 \text{ A m}^2 \text{ kg}^{-1}$). Attributing the difference $\Delta\sigma = \sigma_s - \sigma_{\text{imp}}$ to the graphite, we find an average magnetization σ_c of $23.1 \text{ A m}^2 \text{ kg}^{-1}$, which corresponds to $0.05 \mu_B$ per carbon atom. If the broad component is due to an antiferromagnetic phase, these numbers are 40% greater.

We emphasize that our claim that the magnetization of the graphitic material significantly exceeds anything that can be attributed to the iron-rich phases does not rest on the details of the phase analysis. For example, sample 1.7 contains 44 wt% iron according to the chemical analysis. Of this, only 38% is magnetically ordered. Even if it was all pure iron ($\sigma_s = 220 \text{ A m}^2 \text{ kg}^{-1}$), the magnetization would only be $37 \text{ A m}^2 \text{ kg}^{-1}$, compared with the $52 \text{ A m}^2 \text{ kg}^{-1}$ measured on the same powder. The density, ρ , sets another limit of $38 \text{ A m}^2 \text{ kg}^{-1}$ for this sample, assuming a mixture of graphite and iron. Furthermore, extrapolation of σ_s versus ρ for 18 specimens gives a completely independent value of the average magnetization, 21 (13) $\text{A m}^2 \text{ kg}^{-1}$ (standard deviation is in parentheses) at the density of graphite. A series of acid treatments and density separations failed to eliminate the magnetic moment.

So how does this ferromagnetism come about? One idea is that meteoritic graphite is somehow different from terrestrial graphite, owing perhaps to its mode of formation, chemical doping (Fe, S, P, N...), or the shock of impact. Defects tend to enhance the susceptibility of graphite, and may even lead to superconductivity¹². There are theoretical reports that graphene strips a few nanometres in width show an unusual density of states which leads to local moments at the strip edges which are paramagnetic¹³, or, with suitable stacking, antiferromagnetic¹⁴. But there have been no suggestions that nanographite should be ferromagnetic. Ferromagnetism would not normally be expected in graphite, a semi-metal with Fermi energy $E_F \approx 20$ meV, where the Fermi level falls at a sharp minimum in the density of states¹⁵. Nor is there any indication on our X-ray patterns of intercalation or poor crystallinity on the nanometre scale. The unit cell is about 1% shorter in both a and c directions than usual, with lattice parameters $a_0 = 0.2450(5)$ nm, $c = 0.6713(5)$ nm (experimental error on the last digit is in parentheses), but neither c -axis nor basal-plane reflections are anomalously broad. EDAX analysis of the graphite regions shows no heavy element other than iron at the 0.1 wt% level.

Another idea is that the ferromagnetic inclusions somehow induce a moment in the graphite. Scanning electron micrographs indicate that magnetite and kamacite particles of all sizes down to 50 nm or less are individually embedded in graphite (Fig. 1). Perfectly encapsulated particles are not removed by acid treatment. As magnetite is a half-metal¹⁶, matching the chemical potential of the spin-up and spin-down electrons at the interface leads to complete spin polarization of the adjacent carbon atoms. If the spin splitting of the graphite bands at the Fe_3O_4 interface is a few tenths of an electron volt, the Fermi level there falls in a higher density of states, tending to sustain the spin polarization. The recently discovered gapless spin density mode in graphene sheets⁵

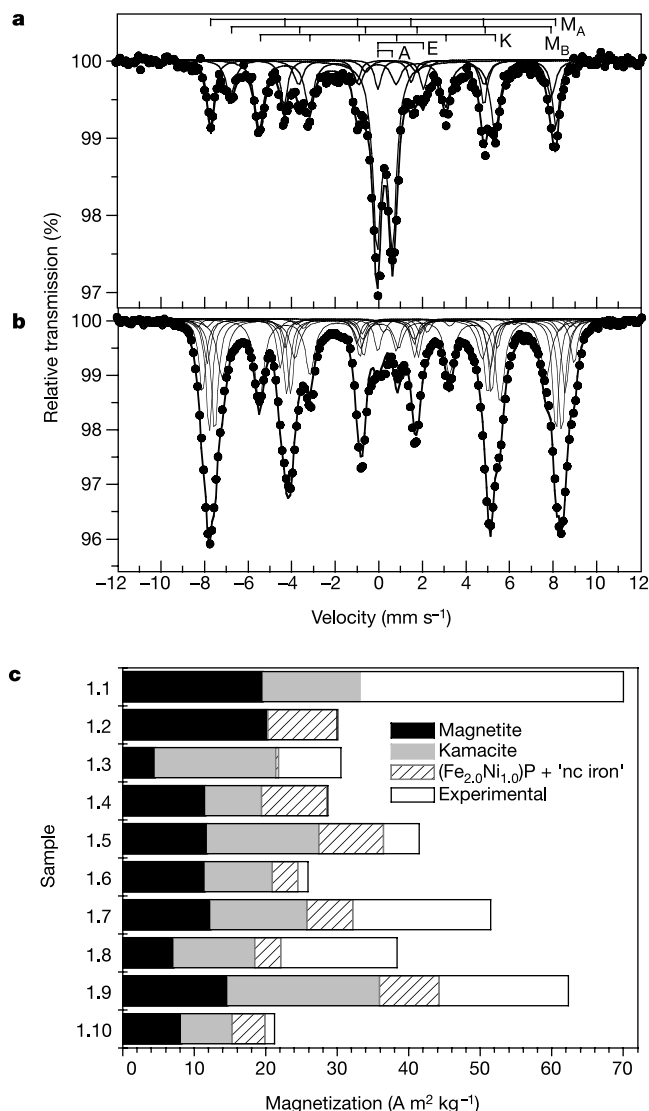


Figure 4 Ferromagnetic phase analysis of samples of graphitic material from the Canyon Diablo meteorite. **a**, **b**, Typical Mössbauer spectra at room temperature (**a**) and at 15 K (**b**). The contributions from the A and B sites of magnetite, kamacite, enstatite and akaganéite are denoted by M_A , M_B , K, E and A, respectively. From the intensity ratio of M_A and M_B , the magnetic sextets with hyperfine fields B_{hf} of 49.0(5) T and 45.8(5) T (experimental error on the last digit is in parentheses), we deduce that the magnetite is oxidized, with composition $\text{Fe}_{2.83}\text{O}_4$. The lattice parameter is $a_0 = 0.8351(5)$ nm. The central paramagnetic doublet with isomer shift $\delta = 0.36(1) \text{ mm s}^{-1}$ relative to $\alpha\text{-Fe}$ at 300 K and quadrupole splitting $\Delta = 0.69(2) \text{ mm s}^{-1}$ is associated with akaganéite. This component is not superparamagnetic in an applied field of 110 mT, as there is no change in the intensity or linewidth of the doublet. The weaker, ferrous doublet (E) with $\delta = 1.17(1) \text{ mm s}^{-1}$ and $\Delta = 2.00(2) \text{ mm s}^{-1}$ is due to iron in M2 sites of enstatite. In most samples, there is the component (K) with $B_{\text{hf}} = 33.5(5)$ T and $\delta = 0.03(1) \text{ mm s}^{-1}$, which is associated with kamacite. A component with $B_{\text{hf}} = 31.0(5)$ T observed in one case is associated with troilite. **c**, Contribution to the magnetization of the ferromagnetic phases (shaded: nanocrystalline iron (nc iron) is probably overestimated, as this phase is likely to be antiferromagnetic). The unshaded residue is associated with graphite.

may greatly enhance the spin susceptibility. Accounting for the magnetization of the graphite in a model where spherical inclusions of radius r_o are dressed by a spin polarization decaying with radius r as $\exp\{-(r-r_o)/\lambda_s\}$, the ratio of the moment of the inclusion and the magnetic carbon shell is approximately $1/3\rho(1+2\rho+2\rho^2)$ where $\rho = \lambda_s/r_o$, and λ_s is the spin decay length. Taking $\sigma_{\text{imp}}/\Delta\sigma = 2.8$, a typical value of $r_o = 50$ nm gives $\lambda_s = 5$ nm. The Fermi wavelength for graphite is 10.6 nm.

The magnetoresistance of the meteoritic graphite shows a positive quadratic variation with applied field B , reaching 2.9% in a field of 2 T at room temperature and 4.8% at 5 K. The in-plane mobility μ in a two-band model with equal concentrations n of electrons and holes is $(\Delta\rho/\rho B^2)^{1/2} \approx 0.1 \text{ m}^2 \text{ V}^{-1} \text{ s}^{-1}$. Taking $n = 7 \times 10^{24} \text{ m}^{-3}$ (ref. 15), the mean free path $\lambda = (\hbar\mu/e)(3\pi^2 n)^{1/3}$ is estimated as 40 nm, and the spin-diffusion length, λ_{sd} , is expected to be an order of magnitude greater. In carbon nanotubes, λ_{sd} is known to exceed 100 nm (ref. 17). Furthermore, there is no resistivity mismatch for a cubic particle embedded in graphite provided that its resistivity lies between the graphite c -axis and a -axis values, as is the case for magnetite. Provided that spin injection can be achieved, the combination of half-metal (magnetite) and semi-metal (graphite) looks promising for spin electronic applications.

Ferromagnetic carbon produced by pyrolysis of organic precursors should be reinvestigated using the present methods, in order to assess if it too is ferromagnetic because of a magnetic proximity effect, as we propose here for meteoritic graphite. The implications of ferromagnetic carbon, whatever its origin, are likely to be wide-reaching—this material could be a zero-gap, high-temperature, ferromagnetic semiconductor. With this new sighting of strongly ferromagnetic meteoritic carbon, we look forward to rapid developments in the area of carbon-based magnetism. □

Received 29 May; accepted 3 September 2002; doi:10.1038/nature01100.

- Esquinazi, P. *et al.* Ferromagnetism in oriented graphite samples. *Phys. Rev. B* **66**, 0244429 (2002).
- Höhne, R. & Esquinazi, P. Can carbon be ferromagnetic? *Adv. Mater.* **14**, 753–756 (2002).
- Makarova, T. L. Magnetism of carbon-based materials. *Studies of High- T_c Superconductivity* (ed. Narlikar, A.) Vol. 44–45 (Nova Science, New York, in the press).
- Makarova, T. L. *et al.* Magnetic carbon. *Nature* **413**, 716–718 (2001).
- Baskaran, G. & Jafari, S. A. Predicting a gapless spin-1 neutral collective mode branch for graphite. *Phys. Rev. Lett.* **89**, 016402 (2002).
- Allemand, P. M. *et al.* Organic molecular soft ferromagnetism in a fullerene C_{60} . *Science* **253**, 301–303 (1991).
- Mrzel, A. *et al.* Ferromagnetism in a cobaltocene-doped fullerene derivative below 19 K due to unpaired spins on fullerene molecules. *Chem. Phys. Lett.* **298**, 329–334 (1998).
- Moehlecke, S., Ho, P. C. & Maple, M. B. Coexistence of superconductivity and ferromagnetism in the graphite-sulfur system. *Phil. Mag. B* **82**, 1335–1347 (2002).
- Murata, K., Ushijima, H., Ueda, H. & Kawaguchi, K. A stable carbon-based organic magnet. *J. Chem. Soc. Chem. Commun.* **7**, 567–569 (1992).
- Grady, M. M. *Catalogue of Meteorites*, 5th Edn 89–90 (Natural History Museum, London, 2000).
- Papike, J. J. (ed.) *Planetary Materials* (Reviews in Mineralogy Vol 35, The Mineralogical Society of America, Washington DC, 1998).
- González, J., Guinea, F. & Vozmediano, M. A. H. Electron-electron interactions in graphene sheets. *Phys. Rev. B* **63**, 134421 (2001).
- Wakabayashi, K., Jujita, M., Ajiki, H. & Sigrist, M. Magnetic properties of nanographites at low temperature. *Physica B* **280**, 388–389 (2000).
- Harigaya, K. The mechanism of magnetism in stacked nanographite: theoretical study. *J. Phys. Condens. Matter* **13**, 1295–1302 (2001).
- Kelly, B. T. *Physics of Graphite* (Applied Science, London, 1981).
- Coey, J. M. D. & Venkatesan, M. Half-metallic ferromagnets; the example of CrO_2 . *J. Appl. Phys.* **91**, 8345–8350 (2002).
- Alphenaar, B. W., Tsukagoshi, K. & Ago, H. Spin electronics using carbon nanotubes. *Physica E* **6**, 848–851 (2000).
- Brett, R. & Higgins, G. T. Cliftonite: A proposed origin and its bearing on the origin of diamonds in meteorites. *Geochim. Cosmochim. Acta* **33**, 1473–1484 (1969).

Acknowledgements We thank M. Viret, F. Guinea and S. Sanvito for discussions. This work was supported by Science Foundation Ireland.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.M.D.C. (e-mail: jcoey@tcd.ie).

Tristable nematic liquid-crystal device using micropatterned surface alignment

Jong-Hyun Kim*, Makoto Yoneya* & Hiroshi Yokoyama*†

* Yokoyama Nano-structured Liquid Crystal Project, ERATO, Japan Science and Technology Corporation, TRC 5-9-9 Tokodai, Tsukuba, Ibaraki 300-2635, Japan

† Nanotechnology Research Institute, National Institute of Advanced Industrial Science and Technology 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan

It has long been appreciated that liquid-crystal (LC) devices in which the LC molecules adopt multiple stable orientations could drastically reduce the power consumption required for high-information-content displays. But for the commonly used nematic LCs, which are intrinsically uniaxial in symmetry, no industrially feasible multi-stable LC device has been realized^{1–7}. Recently we demonstrated how bistability can be robustly engineered into a nematic LC device, by patterning a substrate with an

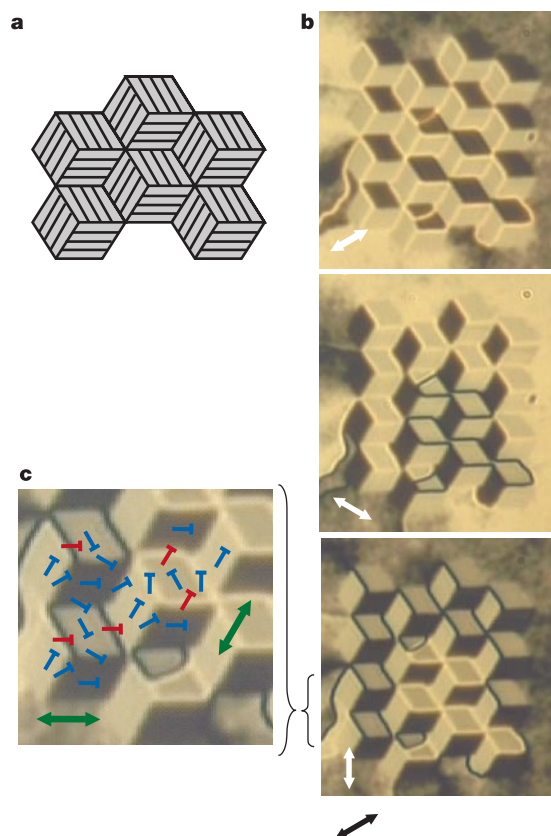


Figure 1 Schematic diagram of the alignment pattern and texture of nematic LC on the surface. **a**, The pattern is an array of identical hexagons, which consist of three elemental parallelograms. Each parallelogram has a shifted local rubbing direction, which is expressed by the solid lines. **b**, The LC texture on the patterned alignment layer under a rotating analyser in the optical polarizing microscope. Outside the patterned area was left as coated. The slight contrast of brightness in the figure indicates that the patterned region is divided globally into two regions of different macroscopic orientations. Black and white arrows indicate the directions of polarizer and analyser. The side of the unit parallelogram is $10 \mu\text{m}$ long. **c**, Orientations of surface directors in individual domains and at boundaries. Green arrows indicate the macroscopic orientations of two regions. Blue 'T' symbols correspond to the director inside the domains and at the boundaries associated with 60° in-plane twist ($\pi/3$ -wall). Red 'T' symbols are the directors at the 120° in-plane twist boundaries ($2\pi/3$ -wall).

orientational chequerboard pattern that enforces orthogonal LC alignment in neighbouring square domains^{8,9}. As a result of the four-fold symmetry of the pattern, the two diagonal axes of the chequerboard become equally stable macroscopic orientations. Here we extend this symmetry approach to obtain a tristable surface-aligned nematic LC. A microscopic pattern exhibiting six-fold symmetry is inscribed on a polyimide surface using the stylus of an atomic force microscope. The hexagonal symmetry of the microscopic orientational domains in turn gives rise to three stable macroscopic LC orientations, which are mutually switchable by an in-plane electric field. The resulting switching mode is surface driven, and hence should be compatible with demanding flexible display applications.

On a fresh surface of a cleaved single crystal such as NaCl or mica, nematic LCs can assume multiply stable orientations, faithfully reflecting the crystallographic axes^{10–12}. Given the microscopic uniformity of the cleavage surface, the ordinary anchoring energy coefficient, which is generally represented by a second rank tensor¹, should be isotropic for surfaces with four-fold or higher-order symmetries. This implies that the multi-stability observed on single-crystal surfaces is a higher-order anchoring effect that makes it susceptible to external contaminants and disturbances leading to the first-order anchoring effect¹³.

Borrowing the concept of symmetry on single crystals while retaining the first-order robustness of our frustrated micropatterning, we devised higher-symmetry surface patterns to realize artificial multi-stability of nematic LCs. As an example of this approach, we demonstrate here tristability on a hexagonally symmetric pattern consisting of locally rubbed parallelograms (Fig. 1). The shape of the parallelogram was designed to have corner angles of 60° and 120° so that three of them, mutually rotated by $\pm 60^\circ$, make a hexagon. The infinitely extended pattern tiled with hexagons (Fig. 1a) has a six-fold rotational symmetry.

For the purpose of examining the aligning characteristics of the

orientational pattern, we fabricated a structure comprising unit domains as large as 10 μm (Fig. 1b). We carried out the nano-rubbing on a polyimide-coated glass slide^{14–17}, the details of which have been described elsewhere^{8,9,17}. In brief, the precursory polyamic acid (SE-150, Nissan Chemical Co.) was spun on the glass slide and baked. For the nano-rubbing, an atomic force microscope (AFM; SPA-500, Seiko Instruments Inc.) was used in the contact mode. The nano-rubbed polyimide surface acquired the ability to align LCs parallel to the surface along the symmetrically scanned direction, giving a zero tilt bias angle¹⁷. A room temperature nematic LC, 4-*n*-pentyl-4'-cyanobiphenyl (5CB), was applied onto the patterned surface to an appropriate thickness without a cover slide.

As the size of the unit domain is large here, LC orientations on individual domains and at the domain boundaries are clearly resolved in the micrographs (Fig. 1b). The contrast changes under a rotating analyser confirm that the alignment inside the unit domain is along the local rubbing direction. The domain boundaries also have a distinct brightness, reflecting the alignment transition from one domain to the next. As indicated by the director symbols in Fig. 1c, two of the three boundaries meeting at a corner are associated with 60° in-plane rotation from one domain to the next ($\pi/3$ -wall), whereas the remaining boundary is associated with 120° rotation ($2\pi/3$ -wall). This combination of walls corresponds to the lowest free-energy configuration that can cover the entire surface without introduction of point defects. Depending on the position of the $2\pi/3$ -wall boundary, there appear three orientationally distinct, yet energetically equivalent, states (Fig. 2). These states are mutually convertible by way of 180° rotation of the director in an appropriate one of the three constituent parallelograms, and are the basis of the present tristability. The global configuration of the director for a given surface orientation is determined in such a way as to minimize orientational distortions described by the Frank elastic energy¹⁸. Consequently, the inhomogeneous LC orientation imposed by the patterned surface relaxes rapidly into a uniform

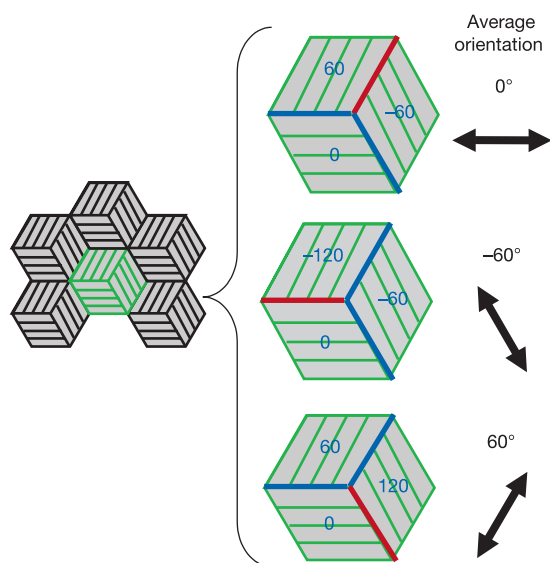


Figure 2 Three possible macroscopic orientations of nematic LC on the hexagonal pattern. The lines inside the pattern indicate the local rubbing direction, in which the LC is locally aligned. The nematic LC and local rubbing are invariant to 180° rotation. The three types of hexagons on the right illustrate the energetically degenerate ground states, which differ from each other in terms of the position of the 120° in-plane twist boundary. Starting from any one of the hexagons, the LC orientations on the remaining two types of hexagons can be obtained by flipping the director by 180° in one of the three unit parallelograms. The macroscopic LC alignment, appearing sufficiently inside the bulk LC, is simply given by averaging the director orientation over the surface, and is 0°, -60° and $+60^\circ$ from top

to bottom. Thick blue and red lines show the $\pi/3$ - and $2\pi/3$ -walls, respectively. Black arrows are the macroscopic LC alignment orientations. The exact match between the macroscopic alignment and either one of the local rubbing directions is valid only when the elastic anisotropy is negligible and/or the inscribed pattern is not chiral. The present hexagonal pattern is chiral in the sense that its mirror image cannot overlap itself. In general, therefore, the macroscopic alignment could be slightly shifted from the local rubbing direction, although the 60° difference among three stable states remains the same owing to its symmetry origin.

bulk alignment, which bears three stable axes depending on the relative arrangement of $\pi/3$ - and $2\pi/3$ -walls. More quantitatively, if we assume the one-constant approximation, the equilibrium configuration of the director azimuth ϕ satisfies the Laplace equation $\Delta\phi = 0$ (ref. 18). It follows that the laterally inhomogeneous component of ϕ with wavevector $\mathbf{k}$ decays exponentially as $\exp(-|\mathbf{k}|z)$ with distance z along the surface normal into the bulk LC. Therefore, the director profile at a point away from the surface by a distance larger than the unit domain size d becomes virtually uniform, with orientation $\phi(z) = \langle \phi(0) \rangle$, where $\langle \rangle$ denotes the average over the surface. This shows that one stable bulk orientation is identical with the local orientation on the unit domain located between the two $\pi/3$ -walls. The remaining two stable states are generated from this configuration by the natural symmetry operation of nematic LCs to flip the surface director by 180° over any particular subset of parallelograms (Fig. 2), giving rise to a 60° shift of $\langle \phi(0) \rangle$.

To confirm the tristable alignment, the hexagonal pattern was scaled down to one-quarter size, giving a side length of $2.4 \mu\text{m}$. As the size of the unit domain approaches the wavelength of visible light, the optical uniformity of the patterned area rapidly improves. This is due to a decrease in the orientational relaxation length $1/|\mathbf{k}|$ and to an increase in the optical diffraction angle beyond the numerical aperture of the observation optics. By combining the patterned substrate with a counter polyimide-coated glass slide that was conventionally rubbed, a $50\text{-}\mu\text{m}$ -thick sandwich-type LC cell was fabricated (Fig. 3A). On the glass substrate three pairs of electrodes, concentrically arranged with 60° shift (Fig. 3A), were fabricated to apply in-plane electric fields. The AFM nano-rubbing was conducted in the central space between electrodes. Nano-rubbing directions were set parallel with the electric field directions. The vacant cell was injected with the nematic LC. A sinusoidal electric field of frequency 1 kHz was used for the electro-optical switching experiments.

The hexagonally patterned area clearly showed three stable states that were mutually switchable by the electric field (Fig. 3B, C). The stable states were found resistant to mechanical disturbances, with no disruption of alignment in prolonged experimentation. This appears reasonable, in that the effective anchoring energy of the tristable cell is reduced by only $1/3$ of a uniform cell, because the switching is associated with the flip of local director on $1/3$ of the unit parallelograms. As the field is increased along the destination direction, the LC in the bulk starts to reorient along the field. Up to a certain threshold field, $\sim 6 \text{ V } \mu\text{m}^{-1}$, the alignment relaxes back to the original state as the field is turned off. When the field strength exceeds the threshold field, the switching begins locally by domain nucleation and growth. Owing to various inhomogeneities associated with LC injection, in particular non-uniform electric and director field distributions on the pattern boundary, the switching is not ideally abrupt, and the field needs to be increased a little before the entire area is switched. Although the three different states switched reversibly between each other, the field strength necessary to drive the a - a' switching was found to be larger than the fields needed for switching b - b' and c - c' (Fig. 3). The difference might reflect the persistent effect of flow alignment during LC injection¹⁹, which took place nearly perpendicular to the a - a' direction. The optical transmission measurements revealed that the contrast ratio was about 5 for all states (Fig. 3C). This ratio was improved to ~ 20 for a tristable device with $1.2\text{-}\mu\text{m}$ unit domain. The angle differences between the stable states were not exactly 60° , but were about 50° symmetrically about the LC injection direction.

The switching behaviour can be readily analysed for the extreme cases of small and large unit domains relative to the anchoring extrapolation length $d_e = K_{22}/W_0$; here, K_{22} and W_0 are the twist elastic constant and anchoring energy, respectively. In the limit of small unit domains $d \ll d_e$, the elastic force overwhelms the surface anchoring, thereby making the alignment microscopically uni-

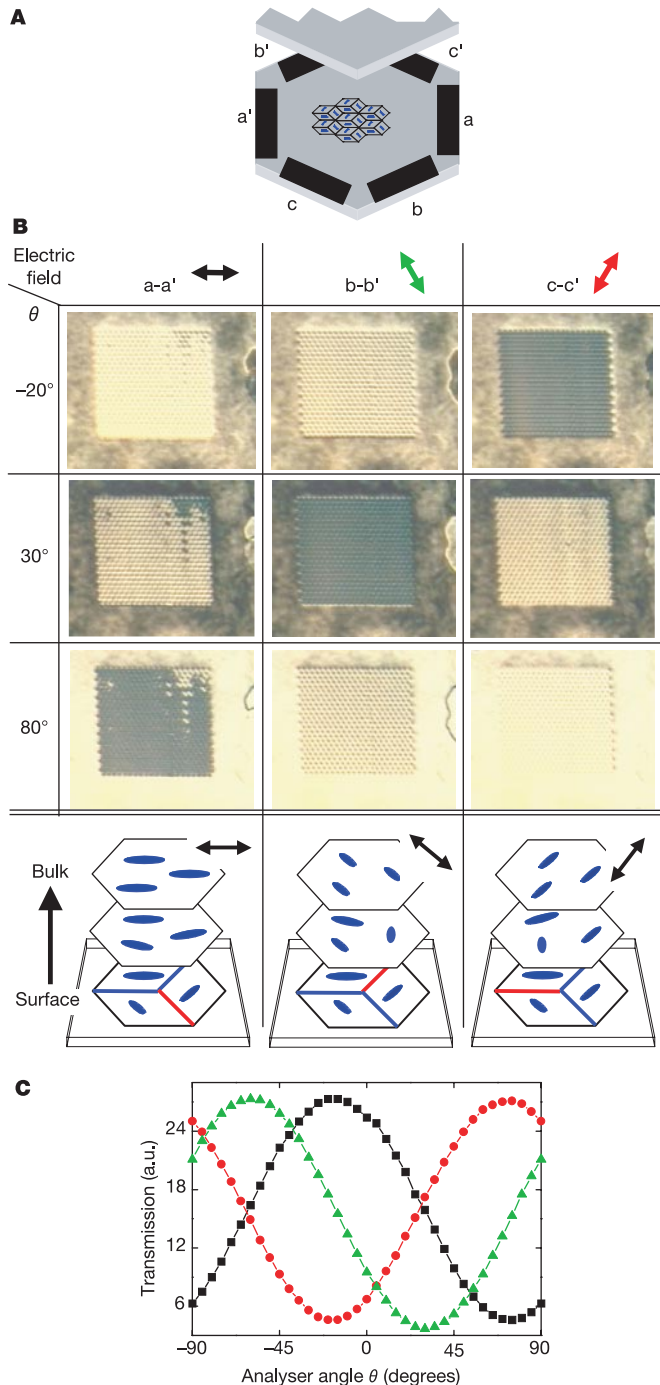


Figure 3 Structure of the LC cell and tristable switching behaviours. **A**, The size of the pattern was about $90 \mu\text{m}$ by $90 \mu\text{m}$. The in-plane electric field was applied along a - a' , b - b' and c - c' for switching. The cell was placed under a polarizing optical microscope for texture observations and for light transmission measurements. Conventional rubbing on the covering substrate and the polarizer were set parallel to each other, and one of the local nano-rubbings was adjusted perpendicular to it. θ is the angle of the analyser from the crossed analyser direction. **B**, The LC textures of the switched states in the absence of electric field; the applied electric field was: $10.0 \text{ V r.m.s. } \mu\text{m}^{-1}$ along a - a' , $7.8 \text{ V r.m.s. } \mu\text{m}^{-1}$ along b - b' and $8.6 \text{ V r.m.s. } \mu\text{m}^{-1}$ along c - c' . The field difference may reflect the inhomogeneity of the cell. The analyser directions were adjusted near the minimum brightness of each state. Most of the area could be switched at lower field strength than these values. Schematic drawings at the bottom are the three-dimensional director profiles for each switched state over the patterned surface into the bulk. As shown in Fig. 2, the red lines are $2\pi/3$ -walls and the blue lines are $\pi/3$ -walls. The black arrows indicate the macroscopic orientation in each switched state. **C**, Transmission as a function of the analyser angle for each switched state in the absence of the electric field. Squares, a - a' direction; triangles, b - b' direction; and circles, c - c' direction. a.u., arbitrary units.

form⁸. For large unit domains $d \gg d_c$, on the other hand, we can treat the system as a uniform planar structure subjected to an in-plane field and obtain the threshold field as $E_{th} = C_M W_0 / 2\sqrt{K_{22}\Delta\epsilon}$ for Rapini-Papoular anchoring¹. Here $\Delta\epsilon$ is the dielectric anisotropy and $C_M = \sqrt{1 + 3\cos^2(2\arctan(\tan^{1/3}|\phi_e - \phi_E|))}$ is a symmetry factor depending on the relative azimuthal angle of the field ϕ_E and the easy direction on the parallelogram ϕ_e . For the tristability $|\phi_e - \phi_E| = 60^\circ$, we obtain $C_M \approx 1.05$, which shows that the threshold field remains only 5% higher than that for the bistability. The observed $E_{th} \approx 6 \text{ V } \mu\text{m}^{-1}$ is consistent with this estimate, given $E_{th} \approx 5 \text{ V } \mu\text{m}^{-1}$ and $W_0 \approx 10^{-4} \text{ J m}^{-2}$ for the bistability⁸. It follows that the angle, $\pm\arctan(\tan^{1/3}|\phi_e - \phi_E|)$, by which the surface director should rotate before the switching, slightly increases from 45° (bistability) to 50.2° (tristability).

As in the ordinary nematic LC devices, cell thickness is a critical parameter determining the electro-optic response. Owing to the patterned nature of the tristable surface, there appears a lower bound for the cell thickness d_t , set by the condition that the surface inhomogeneity should relax sufficiently inside the cell, that is, $d_t \gg d/\sqrt{2\pi}$, to give uniform LC alignment to assure quasi-homogeneous optical and elastic behaviours. Only under this condition can the surface switching be initiated by the in-plane Freedericksz transition in bulk, as required by the switching mechanism⁸. For a domain size of $d \approx 1 \mu\text{m}$, we may adopt a thickness $d_t \approx 5 \mu\text{m}$, indicating that the response speed of the tristable device can at least be comparable to that of existing nematic LC devices. The thickness $d_t \approx 50 \mu\text{m}$ used in the present study is an overly safe choice, solely taken for demonstration purposes.

The present approach to obtaining orientational tristability is believed to be generic; it is based only on the frustration of the periodically aligned LC, and does not rely on any delicate balance of multi-directional forces—it may be extendable to even higher degrees of stability. Consequently, the tristability is not sensitive to the material or to fabrication parameters. We expect that the anisotropic LC properties, coupled with the symmetry of the substrate, will open up a new area of electro-optical applications and interfacial sciences of liquid crystals. □

Received 18 June; accepted 19 September 2002; doi:10.1038/nature01163.

1. Yokoyama, H. *Handbook of Liquid Crystal Research* Ch. 6 (eds Collings, P. J. & Patel, J. S.) (Oxford Univ. Press, New York, 1997).
2. Boyd, G. D., Cheng, J. & Ngo, P. D. T. Liquid-crystal orientational bistability and nematic storage effects. *Appl. Phys. Lett.* **36**, 556–558 (1980).
3. Berreman, D. W. & Heffner, W. R. New bistable liquid-crystal twist cell. *J. Appl. Phys.* **52**, 3032–3039 (1981).
4. Scheffer, T. J. & Nehring, J. A new, highly multiplexable liquid crystal display. *Appl. Phys. Lett.* **22**, 1021–1023 (1984).
5. Yang, K. H. Weak boundary storage effect in homogeneous liquid crystal cells. *Jpn J. Appl. Phys.* **22**, 389–393 (1983).
6. Ong, H., Meyer, R. B. & Hurd, A. J. Multistable orientation in a nematic liquid-crystal cell induced by external-field and interfacial interaction. *J. Appl. Phys.* **55**, 2809–2815 (1984).
7. Barberi, R., Boix, M. & Durand, G. Electrically controlled surface bistability in nematic liquid crystals. *Appl. Phys. Lett.* **55**, 2506–2508 (1989).
8. Kim, J. H., Yoneya, M., Yamamoto, J. & Yokoyama, H. Surface alignment bistability of nematic liquid crystals by orientationally frustrated surface patterns. *Appl. Phys. Lett.* **78**, 3055–3057 (2001).
9. Kim, J. H., Yoneya, M., Yamamoto, J. & Yokoyama, H. Controlling surface alignment on nanoscopically tailored competing domains. *Mol. Cryst. Liq. Cryst.* **367**, 151–158 (2001).
10. Blinov, L. M. & Sonin, A. A. The interaction of nematic liquid crystals with anisotropic substrates. *Mol. Cryst. Liq. Cryst.* **179**, 13–25 (1990).
11. Schuddeboom, P. C. & Jerome, B. Azimuthal anchoring of liquid crystals on surfaces with high symmetry. *Phys. Rev. E* **56**, 4294–4305 (1997).
12. Schuddeboom, P. C. & Jerome, B. Multistable bulk orientation induced by highly symmetric liquid-crystal monolayer. *Europhys. Lett.* **39**, 515–520 (1997).
13. Yoneya, M., Kim, J. H. & Yokoyama, H. Simple model for patterned bidirectional anchoring of nematic liquid crystal and its bistability. *Appl. Phys. Lett.* **80**, 374–376 (2002).
14. Ruetschlim, M., Grutter, P., Funschilling, J. & Guntherodt, H. J. Creation of liquid crystal waveguides with scanning force microscopy. *Science* **265**, 512–514 (1994).
15. Pidduck, A. J., Haslam, S. D., Bryan-Brown, G. P., Bannister, R. & Kiteley, I. D. Control of liquid crystal alignment by polyimide surface modification using atomic force microscopy. *Appl. Phys. Lett.* **71**, 2907–2909 (1997).
16. Rastegar, A., Skarabot, M., Blij, B. & Rasing, Th. Mechanism of liquid crystal alignment on submicron patterned surfaces. *J. Appl. Phys.* **89**, 960–964 (2001).
17. Kim, J. H., Yoneya, M., Yamamoto, J. & Yokoyama, H. Nano-rubbing of a liquid crystal alignment layer by an atomic force microscope: a detailed characterization. *Nanotechnology* **13**, 133–137 (2002).

18. de Gennes, P. G. & Prost, J. *The Physics of Liquid Crystals* (Oxford Univ. Press, New York, 1993).
19. Yamaguchi, R. & Sato, S. Determination of nematic liquid crystal (NLC) orientation by observing NLC droplets on alignment surfaces. *Jpn J. Appl. Phys.* **35**, L117–L119 (1996).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.H.K. (e-mail: kimjh@nanolc.jst.go.jp).

Variability of El Niño/Southern Oscillation activity at millennial timescales during the Holocene epoch

Christopher M. Moy^{*†}, Geoffrey O. Seltzer^{*}, Donald T. Rodbell[‡] & David M. Anderson[§]

^{*} Department of Earth Sciences, 204 Heroy Geology Laboratory, Syracuse University, Syracuse, New York 13244, USA

[‡] Geology Department, Union College, Schenectady, New York 12308, USA

[§] NOAA Paleoclimatology Program, 325 Broadway, and INSTAAR, University of Colorado, Boulder, Colorado 80305, USA

The variability of El Niño/Southern Oscillation (ENSO) during the Holocene epoch, in particular on millennial timescales, is poorly understood. Palaeoclimate studies have documented ENSO variability for selected intervals in the Holocene, but most records are either too short or insufficiently resolved to investigate variability on millennial scales^{1–3}. Here we present a record of sedimentation in Laguna Pallcacocha, southern Ecuador, which is strongly influenced by ENSO variability, and covers the past 12,000 years continuously. We find that changes on a timescale of 2–8 years, which we attribute to warm ENSO events, become more frequent over the Holocene until about 1,200 years ago, and then decline towards the present. Periods of relatively high and low ENSO activity, alternating at a timescale of about 2,000 years, are superimposed on this long-term trend. We attribute the long-term trend to orbitally induced changes in insolation, and suggest internal ENSO dynamics as a possible cause of the millennial variability. However, the millennial oscillation will need to be confirmed in other ENSO proxy records.

A sediment core about 9 m long, retrieved from the lake Laguna Pallcacocha in the southern Ecuadorian Andes, has been used⁴ to investigate Holocene ENSO variability. It was proposed⁴ that the hundreds of light-coloured, inorganic clastic laminae in the sediment core were probably deposited during ENSO-driven episodes of alluvial deposition in the Laguna Pallcacocha drainage basin. This hypothesis was based on the observation that light-coloured laminae deposited in the past 200 yr generally correlated with known El Niño events in instrumental and historical records. The greyscale record, which was used to quantify the distribution of light-coloured laminae, exhibited significant variance in the ENSO band (2–8 yr). However, gaps between adjacent core sections precluded the identification of significant millennial-scale variance. Here we present a new sediment record from Laguna Pallcacocha that is continuous over the past 12,000 yr. We use wavelet analysis to identify a statistically significant millennial oscillation centred on 2,000 yr, and we document broad changes in ENSO- and decadal-band variance over the Holocene that are consistent with other

[†] Present address: Department of Geological and Environmental Sciences, Stanford University, Stanford, California 94305-2115, USA.

ENSO proxy records from the Pacific basin.

During a warm ENSO event, anomalous sea surface temperatures (SSTs) develop off the coast of Ecuador and northern Peru, which initiate widespread convection along the coastal regions of north-western South America. During most of these events, the western Andean slope from 1° to 3° S experiences positive precipitation anomalies⁵. Laguna Pallcacocha is a high-altitude lake (4,200 m above sea level) located 0.5 km east of the western Andean continental divide (2° 46' S, 79° 14' W). The drainage basin has a steep headwall that is mantled in loose debris and debris-flow deposits. The debris flows owe their origin to a combination of weathered andesite and ignimbrite bedrock and sporadic rainfall events. We suggest that during warm ENSO events, convective precipitation (generated from anomalously warm SSTs in the Pacific and orographic uplift along the western flank of the Andes) erodes the

landscape and initiates debris-flow activity within the drainage basin. These events increase stream discharge and sediment load in a single stream that enters Laguna Pallcacocha, from which are deposited the inorganic laminae exposed in our sediment cores. On the basis of comparison with the Quinn El Niño index spanning the past 200 yr, only the moderate-to-strong El Niño events produce clastic layers in the lake. The event frequency that we record is thus lower than the one that captures all El Niño events⁴.

We retrieved two ~8-m cores and two ~0.5-m cores containing the sediment–water interface from the centre of Laguna Pallcacocha in July 1999. At the base of the longer sediment cores are inorganic pink and green glacial silts that grade into 0.75 m of mostly homogeneous dark organic-rich sediment, which further grades into ~6 m of laminated sediment. The laminated section displays hundreds of light-coloured inorganic clastic laminae (~1 mm–1 cm

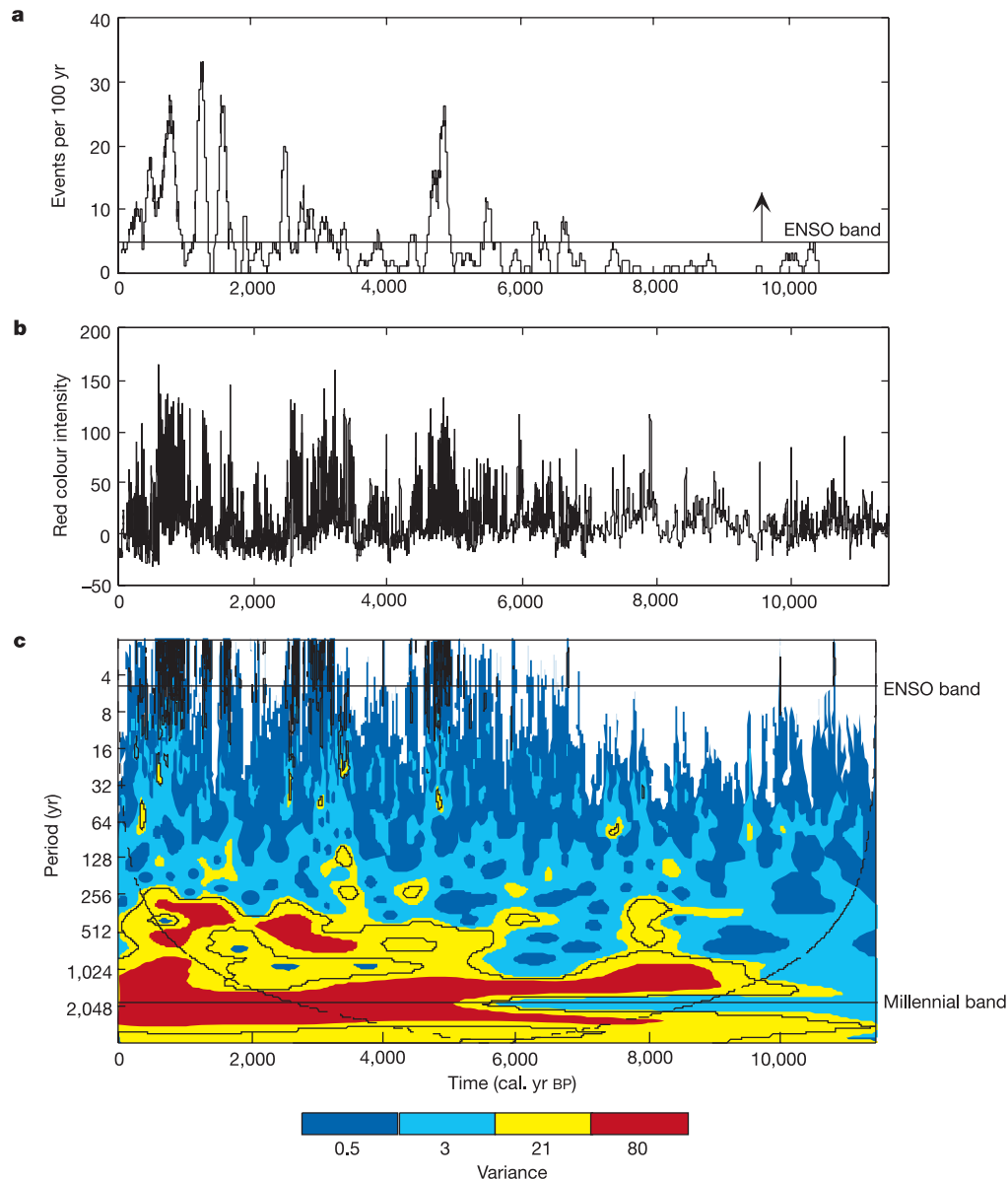


Figure 1 Time series and wavelet power spectrum documenting changes in ENSO variability during the Holocene. **a**, Event time series created using the event model (see Methods), illustrating the number of events in 100-yr overlapping windows. The solid line denotes the minimum number of events in a 100-yr window needed to produce ENSO-band variance (~5). **b**, Most recent 11,500 yr of the time series of red colour intensity. The absolute red colour intensity and the width of the individual laminae do not correspond to the intensity of the ENSO event. **c**, Wavelet power spectrum calculated using the Morlet wavelet

on the time series of red colour intensity (**b**). Variance in the wavelet power spectrum (colour scale) is plotted as a function of both time and period. Yellow and red regions indicate higher degrees of variance, and the black line surrounds regions of variance that exceed the 99.98% confidence level for a red noise process (at 4–8-yr period, the regions of significant variance are shown black rather than outlined). Variance below the dashed line has been reduced owing to the wavelet approaching the end of the finite time series. Horizontal lines indicate average timescale for the ENSO and millennial bands.

thick) interbedded with dark-coloured organic-rich silt. To quantify the distribution of light-coloured laminae through the Holocene, we created a composite record by digitally scanning the surface of the core sections and measuring the reflectance at three wavelengths. We used the intensity of red colour to generate a record of ENSO variability. We then applied two age models that account for the varying sedimentation rate associated with the deposition of the dark organic-rich sediment and the light-coloured laminae, which are deposited within a shorter interval of time (Methods). The event model was used to create an event stratigraphy or event time series illustrating the number of laminae or events in 100-yr overlapping windows (Fig. 1a), while the constant carbon accumulation model was used to create a time series of red colour intensity (Fig. 1b).

Wavelet analysis was used to evaluate the time series of red colour intensity for statistically significant variance. We chose this method because it can be used on non-stationary signals, and depending on what wavelet is used, it can provide excellent time and frequency localization. Wavelet spectra in this study were calculated on the last 11,500 yr of the red colour intensity time series using an algorithm for a continuous wavelet transform with significance testing⁶. We used a Morlet wavelet with a wavenumber of six to calculate the wavelet power spectra and to identify large-scale changes in variance within selected frequency bands through time. The wavelet spectral estimates were tested against the null hypothesis that they are indistinguishable from red noise (99.98% significance level). Red noise characterizes the spectrum found in many climatic time series, and can be represented as a first-order autoregressive process⁷.

The Holocene is characterized by increasing ENSO event frequency towards the present. ENSO variance first becomes statistically significant around 7,000 calibrated years before present (cal. yr BP). The event time series (Fig. 1a) displays an increase in laminae frequency at ~7,000 cal. yr BP. The increase in event frequency is not monotonic, but rather is a gradual pulsing that increases in frequency towards the present. Event frequency culminates at ~1,200 cal. yr BP, and then decreases towards modern times. On the basis of our assumption that Laguna Pallcacocha is probably recording only the moderate and strong events today (and has been doing so throughout the Holocene), we attribute all the variance between 4- and 15-year periods to El Niño activity. The red colour intensity time series displays similar results to the event model (Fig. 1b). High values in the colour intensity time series correspond to the light-coloured laminae. It is important to note that the absolute intensity of red colour and the width of the individual laminae do not correspond to the intensity of the ENSO event. One of the first-order trends seen in the red colour intensity time series is the lower concentration of laminae in the early Holocene, which gradually increases through to the late Holocene. In addition, there are periods of high and low concentrations of light-coloured laminae, which together constitute a millennial-scale oscillation that is persistent throughout the Holocene (Fig. 1b).

A possible complication results from uncertainty in the age model. To investigate possible biases due to our age model, we chose a radiocarbon-dated interval from 2,600 to 3,300 cal. yr BP that is extensively laminated, and artificially stretched the time between radiocarbon dates by a range of factors from 4 to 8. The degree of stretching produces an interval of time far greater than utilizing the minimum 2σ error of the younger date and the maximum 2σ error of the older date in the age model. Variance in the 2–8-yr ENSO band is retained even if the time between radiocarbon-dated intervals is stretched by a factor of eight (ref. 8), indicating that normal errors associated with radiocarbon dating and calibration do not remove the significant variance in the ENSO band in this test. Instead, we found that age model problems can shift the variance to the decadal band. Because of these uncertainties, as well as the fact that the lake is recording only the stronger of the events, we take a conservative stance and attribute all variance between 4 and 15 yr to El Niño variability rather than making

interpretations within the individual ENSO or decadal bands.

A prominent feature of the wavelet spectrum is a millennial-scale oscillation that is coherent throughout the Holocene, but displays less significant variance in the early Holocene. Around 5,000 cal. yr BP the variance shifts from a ~1,500-yr period in the middle/early Holocene to a ~2,000-yr period in the late Holocene (Fig. 1c).

The overall trend exhibited in the Pallcacocha record includes a low concentration of events in the early Holocene, followed by increasing occurrence after 7,000 cal. yr BP, with peak event frequency occurring at ~1,200 cal. yr BP (Fig. 1a). The lack of laminae in the early Holocene can be interpreted in one of two ways. One interpretation is that the ENSO was weak or nonexistent during the early Holocene, and was initiated sometime around 7,000 cal. yr BP. Whereas ENSO may have existed during the early Holocene, it was apparently not strong enough to trigger alluvial deposition within the Laguna Pallcacocha drainage basin. The second possible interpretation is that the drainage basin boundary conditions or lake dynamics were different during the early Holocene, and this precluded the recording of ENSO events in the lake. We favour the first of these explanations, because pollen records from Pallcacocha and a nearby lake reveal no large changes in regional vegetation around the lake at any time during the early Holocene⁹. Similarly, sediment records from alpine lakes in Ecuador and Peru reveal no large changes in organic carbon flux during the Holocene, which indicates that drainage basin parameters remained relatively constant during the past 10,000 yr (refs 10, 11). Our interpretation of a weakened early Holocene ENSO relative to the late Holocene is in agreement with numerous geological proxy records from locations around the Pacific basin^{3,12,13}.

A modelling study that produced results consistent with observations from the Laguna Pallcacocha record is a 15,000-yr run of the Zebiak and Cane ENSO model forced with orbitally induced changes in insolation¹⁴. Such forcing produces reduced ENSO amplitude and frequency during the early Holocene, with a gradual increase in both parameters towards the late Holocene¹⁴. Peak ENSO amplitude occurs around 2,000–1,000 cal. yr BP, and then decreases towards modern times, similar to trends observed here (Fig. 1a). The authors attributed a reduced early Holocene ENSO to increases in boreal summer insolation, which introduce an easterly wind anomaly; this anomaly, through a number of feedbacks, further drives the ENSO system towards a La Niña state by increasing the SST and pressure gradients across the Pacific¹⁴.

The Laguna Pallcacocha record provides evidence for millennial-scale oscillation of ENSO activity during the late Holocene. Two processes known to operate at this timescale are the deposition of ice-rafted detritus in the North Atlantic (Bond events), which have an event pacing of ~1,800 yr during the Holocene¹⁵, and changes in the carbon cycle represented by the residual ¹⁴C record with variance centred on ~2,000 yr (ref. 16). These processes have been used as proxies for North Atlantic deep-water production and solar variability, respectively. We observe that Bond events tend to occur during periods of low ENSO activity immediately following a period of high ENSO activity, which suggests that some link may exist between the two systems. Although the residual ¹⁴C record shares spectral coherence at a period of ~2,000 yr with the time series of red colour intensity, the two time series are not consistently associated with each other during the Holocene. The Zebiak and Cane model has demonstrated that ENSO can display millennial, as well as modern-day, variability even when isolated from the extra-tropics and forced by orbitally induced changes in insolation¹⁷. So although the links with solar and North Atlantic climate warrant further study, internal ENSO dynamics operating independently are a sufficient explanation for the millennial variability that we observe. □

Methods

Acquisition of data

We split the cores and digitally scanned the surfaces with a Geotek linescan camera, which

generates a continuous red, green and blue digital record of the sediment core surface. A single composite record of colour intensity was compiled by sampling the continuous downcore profiles of red, green and blue colour intensity from the core drives at an increment of 0.5 mm. The three colour channels covary ($r^2 = 0.60-0.99$); we chose to use the red colour channel to document the downcore concentration of the laminae in the record because it exhibits slightly higher variance than either the blue or the green channel.

Data analysis

The age model for the new record is based on the same radiocarbon chronology used in ref. 4. The laminae dated by accelerator mass spectrometry ^{14}C of terrestrial microfossils in the record of ref. 4 are distinctive, and could be confidently identified in the new cores. By creating a composite section from overlapping drives, we were able to improve on the original age model. We adopted the constant carbon accumulation model (CCAM) and an event model (EM) to allocate time between dated intervals^{4,8}. The CCAM assumes that the rate of organic carbon deposition has remained nearly constant between radiocarbon-dated intervals through the Holocene, and this continuous sedimentation was punctuated by nearly instantaneous clastic depositional events. The EM is based on the assumption that each of the clastic laminae was deposited during a single ENSO event with a temporal duration of 6 months, and we based this hypothesis on discharge data from the Rio Chira, which display high flow for a period of 6 months during an ENSO event¹⁸. Clastic events were identified in the colour intensity profile as peaks in colour intensity units that exceed the surrounding background laminae by 15 colour intensity units or more⁸. After time was assigned to the clastic laminae, the remaining time was distributed linearly between radiocarbon dates. One advantage of the EM over the CCAM is that it assigns equal magnitude to each of the events, thereby removing any effects that sediment compaction may have on the colour of the light-coloured laminae at the base of the record. Data archived at the NOAA Paleoclimatology Program and available online at: <http://www.ngdc.noaa.gov/paleo/pubs/moy2002>.

Received 15 May; accepted 8 October 2002; doi:10.1038/nature01194.

- Thompson, L. G., Mosley-Thompson, E. & Thompson, P. A. *El Nino: Historical and Paleoclimatic Aspects of the Southern Oscillation* (eds Diaz, H. F. & Markgraf, V.) 295–322 (Cambridge Univ. Press, Cambridge, 1992).
- Cook, E. R., D'Arrigo, R. D., Cole, J. E., Stahle, D. W. & Villalba, R. *El Nino and the Southern Oscillation* (eds Diaz, H. F. & Markgraf, V.) 297–323 (Cambridge Univ. Press, Cambridge, 2000).
- Tudhope, A. W. *et al.* Variability in the El Nino: Southern oscillation through a glacial–interglacial cycle. *Science* **291**, 1511–1517 (2001).
- Rodbell, D. T. *et al.* An ~15,000-year record of El Nino-driven alluviation in southwestern Ecuador. *Science* **283**, 516–520 (1999).
- Vuille, M., Bradley, R. S. & Keimig, F. Climate variability in the Andes of Ecuador and its relation to tropical Pacific and Atlantic sea surface temperature anomalies. *J. Clim.* **13**, 2520–2535 (2000).
- Torrence, C. & Compo, G. P. A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.* **79**, 61–78 (1998).
- Gilman, D. L., Fuglister, F. J. & Mitchell, J. M. On the power spectrum of “red noise”. *J. Atmos. Sci.* **20**, 182–184 (1963).
- Moy, C. M. *A Continuous Record of Late-Quaternary El Nino-Southern Oscillation from the Southern Ecuadorian Andes* Thesis Syracuse Univ. (2000).
- Hansen, B. C. S. *et al.* Late-glacial and Holocene vegetation history from two sites in the western cordillera of southern Ecuador. *Paleogeogr. Paleoclimatol. Paleoecol.* (in the press).
- Rodbell, D. T. The timing of the last deglaciation in Cordillera Oriental, northern Peru, based on glacial geology and lake sedimentology. *Geol. Soc. Am. Bull.* **105**, 923–934 (1993).
- Rodbell, D. T., Bagnato, S., Nebolini, J. C., Seltzer, G. O. & Abbott, M. B. A late glacial–Holocene tephrochronology for glacial lakes in southern Ecuador. *Quat. Res.* **57**, 343–354 (2002).
- Cole, J. Paleoclimate: A slow dance for El Nino. *Science* **291**, 1496–1497 (2001).
- Sandweiss, D. H. *et al.* Variation in Holocene El Nino frequencies: Climate records and cultural consequences in ancient Peru. *Geology* **29**, 603–606 (2001).
- Clement, A. C., Seager, R. & Cane, M. A. Suppression of El Nino during the mid-Holocene by changes in the Earth's orbit. *Paleoceanography* **15**, 731–737 (2000).
- Bond, G. C. *et al.* Mechanisms of Global Change at Millennial Time Scales Geophysical Monograph Series 112 (eds Clark, P. Webb, R. & Keigwin, L. D.) 35–58 (American Geophysical Union, Washington, DC, 1999).
- Stuiver, M., Braziunas, T. F., Becker, B. & Kromer, B. Climatic, solar, oceanic, and geomagnetic influences on Late-Glacial and Holocene atmospheric $^{14}\text{C}/^{12}\text{C}$ change. *Quat. Res.* **35**, 1–24 (1991).
- Clement, A. C. & Cane, M. A. Mechanisms of Global Climate Change at Millennial Time Scales Geophysical Monograph Series 112 (eds Clark, P. U., Webb, R. S. & Keigwin, L. D.) 363–371 (American Geophysical Union, Washington, D.C., 1999).
- Martin, L. *et al.* Southern-Oscillation signal in South-American paleoclimatic data of the last 7000 years. *Quat. Res.* **39**, 338–346 (1993).

Acknowledgements We thank S. Bagnato and J. Turnbull for field assistance, and J. Lewalle for guidance with the wavelet analysis. Discussions with M. Cane, A. Clement and G. Compo significantly improved the manuscript. Funding was provided by the US NSF Earth System History Program (to G.O.S. and D.T.R.), and the Geological Society of America (GSA), the Quaternary Geology and Geomorphology Division of GSA, and the Syracuse University Department of Earth Sciences (to C.M.M.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.M.M. (e-mail: moyc@stanford.edu).

First-principles study of illite–smectite and implications for clay mineral systems

Lars Stixrude & Donald R. Peacor

Department of Geological Sciences, The University of Michigan, Ann Arbor, Michigan 48109, USA

Illite–smectite interstratified clay minerals are ubiquitous in sedimentary basins and they have been linked to the maturation, migration and trapping of hydrocarbons¹, rock cementation², evolution of porewater chemistry during diagenesis³ and the development of pore pressure⁴. But, despite the importance of these clays, their structures are controversial. Two competing models exist, each with profoundly different consequences for the understanding of diagenetic processes: model A views such interstratified clays as a stacking of layers identical to endmember illite and smectite layers, implying discrete and independently formed units (fundamental particles)⁵, whereas model B views the clays as composed of crystallites with a unique structure that maintains coherency over much greater distances, in line with local charge balance about interlayers⁶. Here we use first-principles density-functional theory to explore the energetics and structures of these two models for an illite–smectite interstratified clay mineral with a ratio of 1:1 and a Reichweite parameter of 1. We find that the total energy of model B is 2.3 kJ atom^{−1} mol^{−1} lower than that of model A, and that this energy difference can be traced to structural distortions in model A due to local charge imbalance. The greater stability of model B requires re-evaluation of the evolution of the smectite-to-illite sequence of clay minerals, including the nature of coexisting species, stability relations, growth mechanisms and the model of fundamental particles.

Illite and smectite both consist of aluminosilicate layers alternating with interlayers. In illite the charge of non-exchangeable interlayer cations is balanced by a greater net negative charge on the layers due largely to the substitution of Al for Si in the tetrahedral sheets. The mineral rectorite is the regularly interstratified sequence of illite- and smectite-like layers with a ratio of 1:1 and a Reichweite (R) ordering parameter R1. The illite- and smectite-like components of rectorite may be viewed as being centred either on the layers or on the interlayers (Fig. 1). The former view corresponds to model A: as in pure illite or smectite, the two tetrahedral sheets of each layer have the same composition; each interlayer is identical and is adjacent to a high-charge and a low-charge layer. In model B interstratification is viewed as being interlayer centred: each layer is identical and contains an Al-rich and an Al-poor tetrahedral sheet; up-down alternation of these layers results in two distinct interlayers, one adjacent to low-charge tetrahedral sheets and the other to high-charge tetrahedral sheets. Model B is consistent with alternating K-rich and K-poor interlayers, as implied by X-ray diffraction⁷.

Small crystal sizes, complex composition, and turbostratic stacking (disordered rotations about the layer normal) in illite–smectite clays prevent complete experimental specification of the structure. However, advances in solid-state theory and computing power allow the exploration of even complex crystal structures such as clays. To evaluate the energetics of illite–smectite clays, and to test models A and B, we performed first-principles quantum mechanical calculations based on density-functional theory in the local density approximation. We considered the system $\text{K}_x\text{Al}_2(\text{Al}_x\text{Si}_{4-x})\text{O}_{10}(\text{OH})_2$, where $x = 1$ (muscovite) and $x = 0$ (pyrophyllite) are endmembers of the illite–smectite series. We determined the total energies of fully

relaxed structures of models A and B for $x = 0.5$, which corresponds to a 1:1 R1 illite–smectite (I–S) composition (Fig. 1). For each model, the cell shape and internal atomic coordinates were systematically adjusted until the energy was minimized and the stress was hydrostatic. Energy minimization was performed at constant volume, which we assumed to be that of muscovite ($V = 943 \text{ \AA}^3$ per 82 atoms). All computations were performed with the plane-wave pseudopotential method using the Vienna Ab Initio Simulation Package (VASP)⁸. This method has been used successfully to study clays⁹.

We find that model A has a total energy that is substantially higher than that of model B. The difference in energy (24 meV atom^{-1} , or $2.3 \text{ kJ atom}^{-1} \text{ mol}^{-1}$) is comparable to the thermal energy at ambient temperature. For comparison, the quartz–cristobalite enthalpy difference at ambient conditions is more than a factor of two smaller ($1 \text{ kJ atom}^{-1} \text{ mol}^{-1}$) (ref. 10). The energy difference can be understood in terms of the two model structures: whereas in model B the excess charge of the interlayer K is balanced locally by charge-deficient tetrahedral sheets, in model A charge deficiency is disposed asymmetrically about the interlayer, allowing for only partial local balance of the interlayer charge.

The computations provide additional insight into the structural energetics of the two models. In model A, the asymmetric charge about the interlayer leads to a displacement of the K ion towards the charge-deficient layer by $0.7 \text{ \AA}$. The proximity of the K ion causes substantial distortions of the coordination polyhedra, which contribute to the energetic instability of model A: the Al tetrahedron shows a variance of bond angles (units: degrees squared) (ref. 11) $\sigma = 46$, and three of the Si tetrahedra have $\sigma > 29$. In contrast, tetrahedra in model B show $2 < \sigma < 10$, within the range of values found experimentally in crystal-structure refinements of clays and micas (0.2–14.2) (ref. 12). As expected for dioctahedral phyllosilicates, OH bonds in both models are inclined towards the vacant site

in the octahedral sheet, forming an angle (δ) with respect to the layer normal. In model B, the OH bonds about the K interlayer show $\delta = 69\text{--}75^\circ$, in agreement with experimental measurements in muscovite^{13,14}. In contrast, model A shows much larger inclinations: the OH bonds nearest to the K ion show $\delta = 97\text{--}99^\circ$, that is, the hydrogen ions are forced below the plane of oxygen ions to which they are bonded.

The fact that differences in energy and structure between the two models can be understood in terms of local charge balance is important for the applicability of our results to natural illite–smectite. Layers of illite–smectite are known to stack in a variety of ways, including turbostratically. Different stacking sequences must result in small adjustments in local structure. However, such adjustments and the resultant differences in energy are trivial compared with the major differences between the tested models, and do not affect the conclusions of this study. In natural samples, the difference in charge between low- and high-charge elements is reduced as compared with our idealized structures based on endmembers. Differences in energy and structure between models A and B are expected to be reduced proportionately, but to remain large to the extent that low-charge and high-charge elements remain distinct.

The structure of model B is significant because it shows that the energetically favourable state of R1 I–S is not a simple mixture of endmember layers of illite and smectite. The structural components of model B, including the 2:1 layers (tetrahedral–octahedral–tetrahedral, TOT), are not found in either endmember. This has important implications for the thermodynamic properties of illite–smectite clays. Because these are difficult to obtain experimentally at low temperatures of formation, there have been many previous attempts to model thermodynamic properties¹⁵. These models did not consider the unique low-energy ordered state of R1 I–S represented by model B. The difference in energy between model B and a nanoscale mechanical mixture of illite and smectite layers (model A) is substantial and must be considered in future modelling of thermodynamic properties.

The unique structure of R1 I–S shows that it is a relatively stable phase of low potential energy, as demonstrated also for smectite and illite by their common occurrence. Our results imply that the

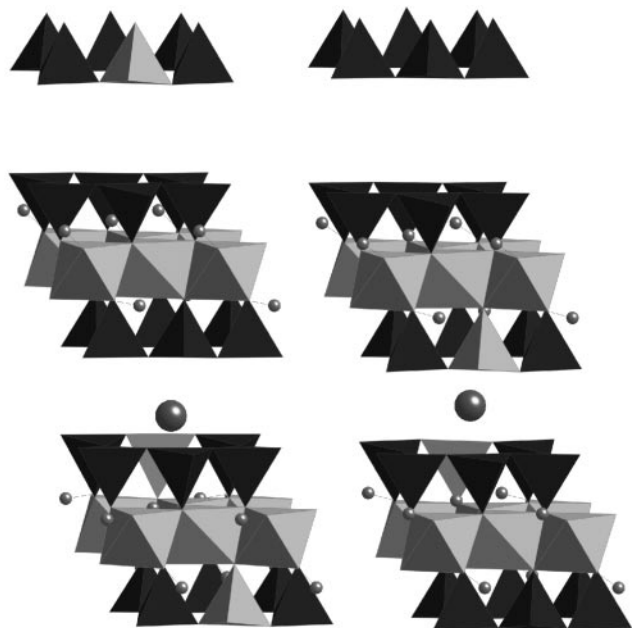


Figure 1 Fully relaxed computed structures of model A (left) and model B (right) 1:1 illite–smectite. Dark tetrahedra, Si; light tetrahedra, Al; large dark spheres, K; small dark spheres, H. Al octahedra are also shown. The repeat unit in the c -direction in each case contains one occupied and one unoccupied interlayer. For all calculations, we assumed an energy cut-off of 500 eV and a $2 \times 2 \times 2$ Monkhorst–Pack²⁹ k -point mesh. Computed lattice parameters for model A: $a = 5.168$, $b = 8.947$, $c = 20.808 \text{ \AA}$, $\alpha = 89.97^\circ$, $\beta = 101.40^\circ$, $\gamma = 89.96^\circ$; and model B: $a = 5.130$, $b = 8.868$, $c = 21.163 \text{ \AA}$, $\alpha = 89.82^\circ$, $\beta = 101.61^\circ$, $\gamma = 89.96^\circ$, are within 2% of those measured for a natural K-rich rectorite without interlayer water³⁰.

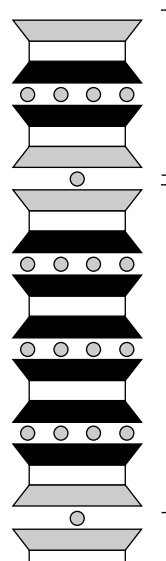


Figure 2 Schematic of illite-rich illite–smectite structure showing interlayer cations (spheres) and 2:1 layers. Shading of the tetrahedral sheets indicates relative concentration of tetrahedral Al, with the darkest shading corresponding to the highest concentrations (greatest charge deficiency). Two illite-like units, bounded by cleavage planes, one 20-Å thick and one 40-Å thick, are indicated by the brackets.

commonly observed assemblages in the illite–smectite series should be pure smectite, smectite + R1 I–S, pure R1 I–S, R1 I–S + illite, and illite. The series is not continuous, but it has gaps between smectite and R1 I–S, and between illite-rich illite–smectite and illite, such gaps being analogous to immiscibility gaps. Indeed, recent transmission electron microscope (TEM) studies¹⁶ demonstrated the occurrence of these assemblages. TEM data show a large gap between smectite and R1 I–S, and a smaller gap between a range of illite-rich illite–smectite and illite. Figure 2 shows the structure relations for illite-rich illite–smectite. The Al–Si ordering implied by our model verifies that it can be viewed as interstratified packets of illite and R1 I–S (ref. 17), not illite and smectite layers. This does not mean that complete sequences of randomly ordered smectite-like and illite-like layers cannot occur. Such interstratified clays may form where crystallization favours extreme disorder. Our results do imply, however, that much if not most of ‘mixed-layered illite–smectite’ involves phases for which ordered R1 I–S is the critical, unique phase, analogous to corrensite in the chlorite–smectite (saponite) system¹⁸.

This model at first appears incompatible with X-ray diffraction (XRD) data, which are consistent with a continuous series of interstratified illite–smectite (ref. 19) as expressed in Reichweite nomenclature²⁰, and which provides support to fundamental particle theory⁵. We tentatively suggest that the differences in interpretations of XRD patterns and TEM observations are caused by differences in sample preparation. TEM analysis involves ion-milled samples that retain their original texture, whereas XRD analysis generally involves mechanical and chemical treatment that causes separation of layers, especially along low-charge interlayers, which are reconstituted in altered sequences in samples for which interparticle diffraction determines characteristics of XRD patterns²¹.

Fundamental particle theory⁵ requires that illite–smectite comprise illite and smectite layers and is thus incompatible with the results of this study. The theory assumes that R1 I–S comprises independent elementary particles of illite (20-Å-thick units centred on an illite interlayer) and smectite (10-Å-thick elementary smectite particles). Illite-rich illite–smectite is modelled as equivalent to independent fundamental particles comprising coherently related illite layers. Interfaces between particles in aggregates are modelled as being smectite-like and, through interparticle X-ray diffraction, as giving rise to sequences that behave like illite–smectite (ref. 21). Fundamental particles are theorized to exist not only in samples prepared for XRD, but also in original geologic samples. This model has been criticized²² on the basis of TEM observations of intact samples. Our results show that the 2:1 layers in a sequence of illite-rich illite–smectite are not independent (Fig. 2). The relative proportions of Al and Si in any one layer are directly related to those in adjacent layers. Moreover, such layer sequences are compatible with the data for thickness of separates on which the fundamental particle theory is based⁵. Cleavage across low-charge interlayers must produce separate illite-like particles (Fig. 2), but with interfaces that are truly low charge (Al-poor tetrahedral sheets), consistent with observed relations^{6,23} and inconsistent with the smectite-like behaviour attributed to fundamental particles. We suggest, therefore, that fundamental particle theory is based on an invalid crystal-chemical model and incorrectly implies equality of states of clay separates and crystallites in unprocessed rocks.

Several mechanisms for the formation of illite–smectite during diagenesis of pelitic sediments have been proposed, but they can be approximated as two types, layer-by-layer replacement and dissolution/neocrystallization²⁴. Our results favour dissolution and neocrystallization^{25,26} because the differences in Al–Si ordering between rectorite and illite or smectite are incompatible with transformations by means of solid-state diffusion of Al and Si at the low transformation temperatures observed in Nature²⁵. Previous work, based on XRD patterns, has indicated layer-by-layer replacement of

smectite by illite, for example in Gulf Coast mudstones, which show a continuous series of mixed-layered illite–smectite with increasing depth by XRD²⁷. The layer-by-layer mechanism, however, is incompatible with the presence of discrete phases separated by gaps.

The validity of theories of crystal growth is based on observations of the sizes and shapes of crystals in evolving systems. The most recent theories of growth of illite–smectite have been tested assuming that particle thicknesses (as measured in separates by XRD) are fundamental particles; that is, they consist only of the sequences of illite layers obtained in separates²⁸. As shown in Fig. 2, however, layers that are separated by a low-charge interlayer (smectite-like) are not independent, and thus a layer added during growth is not a randomly determined illite or smectite layer. Adjacent layers have structures that are mutually dependent, as is true of any crystal structure and defined in the case of illite–smectite by the distribution of Al in the tetrahedral sheets. We view the common lack of coherency (turbostratic stacking) that occurs across low-charge interlayers as a defect in a continuous crystal, albeit a defect that is the rule rather than the exception; that is, packets of layers are analogous to McEwan crystallites⁶. No theory of crystal growth can be applied successfully to measures of crystal size that are not equivalent to actual growth units. The kinds of complete layer sequences shown in Fig. 2 are inferred to represent growth units, in contrast to units consisting only of the illite-like components in separates as interpreted by fundamental particle theory. We consider that determinations of growth mechanisms in natural systems that assume the existence of fundamental particles in clay separates are therefore in error. □

Received 21 May; accepted 19 September 2002; doi:10.1038/nature01155.

- Weaver, C. E. Possible uses of clay minerals in search for oil. *Bull. Am. Assoc. Petrol. Geol.* **44**, 1505–1518 (1960).
- Boles, J. R. & Franks, S. G. Clay diagenesis in Wilcox sandstones of southwest Texas—implications of smectite diagenesis on sandstone cementation. *J. Sedim. Petrol.* **49**, 55–70 (1979).
- Brown, K. M., Saffer, D. M. & Bekins, B. A. Smectite diagenesis, pore-water freshening, and fluid flow at the toe of the Nankai wedge. *Earth Planet. Sci. Lett.* **194**, 97–109 (2001).
- Bethke, C. M. Inverse hydrologic analysis of the distribution and origin of Gulf Coast-type geopressed zones. *J. Geophys. Res.* **91**, 6535–6545 (1986).
- Nadeau, P. H., Wilson, M. J., McHardy, W. J. & Tait, J. M. Interstratified clays as fundamental particles. *Science* **255**, 923–925 (1984).
- Altaner, S. P., Weiss, C. A. & Kirkpatrick, R. J. Evidence from ²⁹Si NMR for the structure of mixed layer illite–smectite clay minerals. *Nature* **331**, 699–702 (1988).
- Brown, G. Crystal-structures of clay-minerals and related phyllosilicates. *Phil. Trans. R. Soc. Lond. A* **311**, 221–240 (1984).
- Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
- Stixrude, L. Talc under tension and compression: spinodal instability and structure at high pressure. *J. Geophys. Res.* (in the press).
- Berman, R. G. Internally-consistent thermodynamic data for minerals in the system Na₂O–K₂O–CaO–MgO–FeO–Fe₂O₃–Al₂O₃–SiO₂–TiO₂–H₂O–CO₂. *J. Petrol.* **29**, 445–522 (1988).
- Robinson, K., Gibbs, G. V. & Ribbe, P. H. Quadratic elongation—quantitative measure of distortion in coordination polyhedra. *Science* **172**, 567–570 (1971).
- Smyth, J. R. & Bish, D. L. *Crystal Structures and Cation Sites of the Rock-Forming Minerals* (Allen and Unwin, Boston, 1988).
- Rothbauer, R. Untersuchung eines 2M1-Muskovits mit Neutronenstrahlen. *Neues Jb. Mineral. Mh.* **4**, 143–154 (1971).
- Rossmann, G. R. Spectroscopy of micas. *Rev. Mineral.* **13**, 145–181 (1984).
- Essene, E. J. & Peacor, D. R. Clay mineral thermometry—A critical perspective. *Clays Clay Minerals* **43**, 540–553 (1995).
- Dong, H., Peacor, D. R. & Freed, R. L. Phase relations among smectite, R1 illite-smectite, and illite. *Am. Mineral.* **82**, 379–391 (1997).
- Schroeder, P. A. & Irby, R. Detailed X-ray diffraction characterization of illite-smectite from an Ordovician K-bentonite, Walker County, Georgia, USA. *Clay Minerals* **33**, 671–674 (1998).
- Shau, Y. H., Peacor, D. R. & Essene, E. J. Corrensite and mixed-layer chlorite/corrensite in metabasalt from northern Taiwan: TEM/AEM, EMPA, XRD, and optical studies. *Contrib. Mineral. Petrol.* **105**, 123–142 (1990).
- Srodon, J., Morgan, D. J., Eslinger, E. V., Eberl, D. D. & Karlinger, M. A. Chemistry of illite/smectite and end-member illite. *Clays Clay Minerals* **34**, 368–378 (1986).
- Srodon, J. Nature of mixed-layer clays and mechanisms of their formation and alteration. *Annu. Rev. Earth Planet. Sci.* **27**, 19–53 (1999).
- Nadeau, P. H., Wilson, M. J., McHardy, W. J. & Tait, J. M. Interparticle diffraction: a new concept for interstratified clays. *Clay Minerals* **19**, 757–769 (1984).
- Kasama, T., Murakami, T., Kohyama, N. & Watanabe, T. Experimental mixtures of smectite and rectorite: Re-investigation of “fundamental particles” and “interparticle diffraction”. *Am. Mineral.* **86**, 105–114 (2001).
- Jakobsen, H. J., Nielsen, N. C. & Lindgreen, H. Sequences of charged sheets in rectorite. *Am. Mineral.* **80**, 247–252 (1995).

24. Altaner, S. P. & Ylagan, R. F. Comparison of structural models of mixed-layer illite-smectite and reaction mechanisms of smectite illitization. *Clays Clay Minerals* **45**, 517–533 (1997).
25. Ahn, J. H. & Peacor, D. R. A transmission and analytical electron microscopic study of the smectite to illite transition. *Clays Clay Minerals* **34**, 165–179 (1986).
26. Nadeau, P. H., Wilson, M. J., McHardy, W. J. & Tait, J. M. The conversion of smectite to illite during diagenesis: evidence from some illitic clays from bentonites and sandstones. *Mineral. Mag.* **49**, 393–400 (1985).
27. Hower, J., Eslinger, E. V., Hower, M. E. & Perry, E. A. Mechanism of burial metamorphism of argillaceous sediments: Mineralogical and chemical evidence. *Geol. Soc. Am. Bull.* **87**, 725–737 (1976).
28. Eberl, D. D., Drits, V. A. & Srodon, J. Deducing growth mechanisms for minerals from the shapes of crystal size distributions. *Am. J. Sci.* **298**, 499–533 (1998).
29. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **13**, 5188–5192 (1976).
30. Benincasa, E., Brigatti, M. F., Medici, L. & Poppi, L. K-rich rectorite from kaolinized micaschist of the Sesia-Lanzo Zone, Italy. *Clay Minerals* **36**, 421–433 (2001).

Acknowledgements This work was supported by the US National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to L.S. (e-mail: stixrude@umich.edu).

Synchronization of animal population dynamics by large-scale climate

Eric Post* & Mads C. Forchhammer†

* Department of Biology, The Pennsylvania State University, 208 Mueller Lab, University Park, Pennsylvania 16802, USA

† Department of Population Ecology, Zoological Institute, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen, Denmark

The hypothesis that animal population dynamics may be synchronized by climate¹ is highly relevant in the context of climate change because it suggests that several populations might respond simultaneously to climatic trends if their dynamics are entrained by environmental correlation. The dynamics of many species throughout the Northern Hemisphere are influenced by a single large-scale climate system, the North Atlantic Oscillation (NAO)^{2,3}, which exerts highly correlated regional effects on local weather⁴. But efforts to attribute synchronous fluctuations of contiguous populations to large-scale climate are confounded by the synchronizing influences of dispersal or trophic interactions⁵. Here we report that the dynamics of caribou and musk oxen on opposite coasts of Greenland show spatial synchrony among populations of both species that correlates with the NAO index. Our analysis shows that the NAO has an influence in the high degree of cross-species synchrony between pairs of caribou and musk oxen populations separated by a minimum of 1,000 km of inland ice. The vast distances, and complete physical and ecological separation of these species, rule out spatial coupling by dispersal or interaction. These results indicate that animal populations of different species may respond synchronously to global climate change over large regions.

Despite widespread evidence that several species' populations respond to large-scale climatic fluctuation^{2,3}, attributing spatial synchrony in animal population dynamics to environmental forcing is problematic⁶. The correlated dynamics of populations of Soay sheep (*Ovis aries*) on separate islands⁷ is strongly indicative of spatial coupling by climate, as is the segregation of the structural dynamics of Canada lynx (*Lynx canadensis*) populations into regions experiencing similar climatic regimes⁸. But the high degree of gene flow among the lynx populations indicates that there is also a high degree of dispersal among them⁹. The pervasive influence of the NAO¹⁰ in the population dynamics of vertebrates throughout

the Northern Hemisphere^{2,8,11–13}, including Soay sheep^{13–15} and Canada lynx⁸, suggests that the NAO has the potential to contribute to spatial synchrony across broad geographic scales in these and other populations¹⁶. The challenge therefore lies in identifying systems in which the climatic signal in population synchrony is not obscured by dispersal or trophic interactions.

We have analysed continental-scale data on the long-term dynamics of caribou (*Rangifer tarandus*) and musk oxen (*Ovibos moschatus*) in Greenland, where a continent-wide ice sheet separates the two species physically and ecologically (ref. 17 and Fig. 1a). Caribou are indigenous to west Greenland but do not inhabit northeast Greenland, where musk oxen are indigenous (Fig. 1a). Although time series data on population dynamics exist for seven caribou and six musk oxen populations in Greenland, our analysis focuses primarily on six caribou and five musk oxen populations whose dynamics are influenced by the NAO (Methods). The physical isolation of these two species, coupled with their lack of interspecific competition for food or shared predators in Greenland, greatly simplifies the analysis of climate-induced synchrony. In addition, evidence indicates that the dynamics of populations of

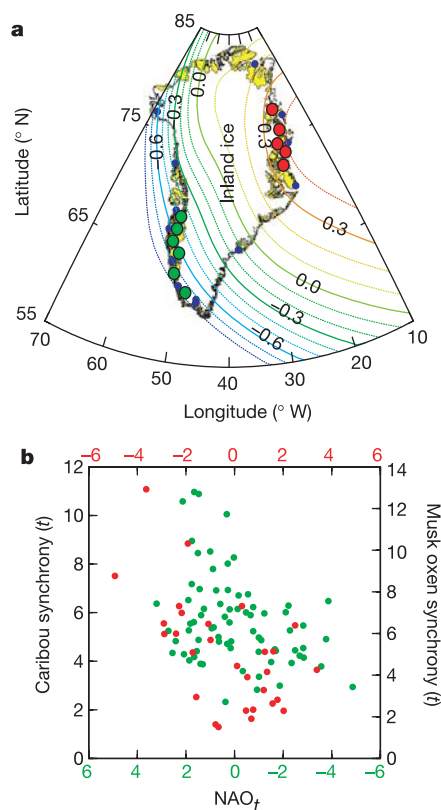


Figure 1 Correlations between the NAO and local weather and population synchrony in Greenland. **a**, Spatial gradient in the correlation between local winter temperature and the winter NAO index across Greenland. Contours (interval 0.1) denote correlations between the NAO winter index and average winter (December to March) temperature from 12 weather stations (blue dots) over the period 1967 to 1995. Correlative contours were constructed as described²⁹ using linear interpolation³⁰. Red and green circles show the location of the respective musk oxen and caribou populations used in our analyses. **b**, Correlations between the winter NAO index and synchrony (Methods) across populations of musk oxen (red; $R^2 = 0.45$, $P < 0.001$) and caribou (green; $R^2 = 0.43$, $P < 0.001$) in northeast and west Greenland, respectively. Inclusion of the two populations that are not influenced by the NAO resulted in poorer correlations for both species (Methods). The top x-axis applies to the correlation with musk oxen synchrony, the bottom x-axis applies to the correlation with caribou synchrony; axes are reversed to show the same gradient from cold (left) to warm (right).

both species in Greenland are influenced by the effect of the NAO on local winter temperatures¹⁸; notably, these temperatures correlate with the NAO index along a spatial gradient across Greenland, from positive correlation on the east coast to negative correlation on the west coast (Fig. 1a). Consequently, any potentially synchronizing influence of the NAO on the dynamics of caribou and musk oxen populations in Greenland should be apparent in a similar spatial dichotomy from east to west.

Among those populations in each species whose dynamics are influenced by the NAO, interannual variation in the degree of spatial synchrony correlates significantly with interannual fluctuations in the winter NAO index (Fig. 1b). Inclusion of the two populations that are not influenced by the NAO weakened this relationship for both species (Methods). Thus, the climatic signal in population synchrony is clearest when examining only those populations that are influenced by climate¹. The dichotomous influence of the NAO on population synchrony in each species (positive for caribou and negative for musk oxen; Fig. 1b) is in agreement with the dichotomous influence of the NAO on local winter temperatures on opposite coasts of Greenland (Fig. 1a) and suggests that population synchrony in both species is highest after cold winters and lowest

after warm winters. Cold winters might synchronize population fluctuations in each species through increased juvenile mortality resulting, for example, from ice-crust formation^{19,20}. Alternatively, increased spatial synchrony of plant phenology after cold winters^{13,31} might synchronize offspring production in both species, because timing and synchrony of parturition in arctic ungulates are highly correlated with plant phenology^{21,22}.

Population synchrony is also evident in the correlated dynamics among populations of both species, and scatter plots reveal a decline in the degree of pairwise correlation between populations with increasing distance separating them (Fig. 2a). A high degree of spatial correlation is also apparent in local winter temperatures measured at stations along each coast (mean cross-correlation between weather stations, west coast $r = 0.94$; east coast $r = 0.69$). Although these correlations, and those in Fig. 1b, suggest that synchrony among populations in each species may relate to spatial auto-covariation in local winter weather driven by the NAO, we cannot rule out the potential synchronizing influence of dispersal. This seems particularly relevant to the dynamics of caribou, which undertake long-distance migrations; indeed, inclusion of the single caribou population that is not influenced by the NAO reveals fairly high synchrony with some of the other caribou populations (Fig. 2a).

Clearer support for the hypothesis of spatial coupling of these populations by climatic fluctuation is evident, however, in an analysis of cross-species synchrony. Calculation of all possible pairs of cross-species correlations shows that the dynamics of caribou and musk oxen are highly synchronized over a range of considerable distances, from a minimum of 1,000 km up to about 1,700 km (Fig. 2a). At distances beyond 1,700 km, cross-species pairs of populations are increasingly out of phase with each other (Fig. 2a). The largely similar spatial pattern of correlation in local winter temperatures measured at weather stations on opposite coasts of Greenland, including negative covariation at distances greater than 1,700 km, suggests that this population synchrony is attributable to spatio-temporal climatic variation (Fig. 2b).

The results shown in Fig. 2a present a conundrum: the cross-species population correlations of many pairs are higher than the within-species population correlations. We suggest that this may reflect the combined, and perhaps opposing, influences of dispersal and climate on population synchrony in each species. Both caribou and musk oxen migrate seasonally and may disperse under certain climatic conditions^{19,20}. The direct climatic influence on population synchrony, operating perhaps through winter mortality, might be distinct from the climatic conditions that trigger long-range dispersal because such movements tend to occur when climatic conditions are favourable for population increases²⁰. In addition, intra-species population synchrony might reflect the influence of spatially autocorrelated noise on a more local scale overlaid by that of dispersal and large-scale noise⁶, that is, the NAO. By contrast, cross-species correlation is not, in this example, muddled by local noise or dispersal among populations; nor is it obscured by shared predators or competition because all of the populations are completely separated geographically.

To investigate further the potential synchronizing influence of large-scale climate in cross-species correlations, we applied a technique analogous to the log-response ratio commonly used in meta-analysis²³. This analysis revealed a highly significant positive correlation between cross-species population synchrony and the magnitude of the effect of the NAO on cross-species pairs of populations (Fig. 3). As Moran's¹ analyses and reiterations of it have shown²⁴, the degree of pairwise correlation between any two populations synchronized by climate should scale with the degree of correlation between local weather variables influencing the dynamics of those populations. Because the NAO has opposing influences on the local weather in the regions inhabited by the focal species studied here, relating the degree of pairwise population synchrony to

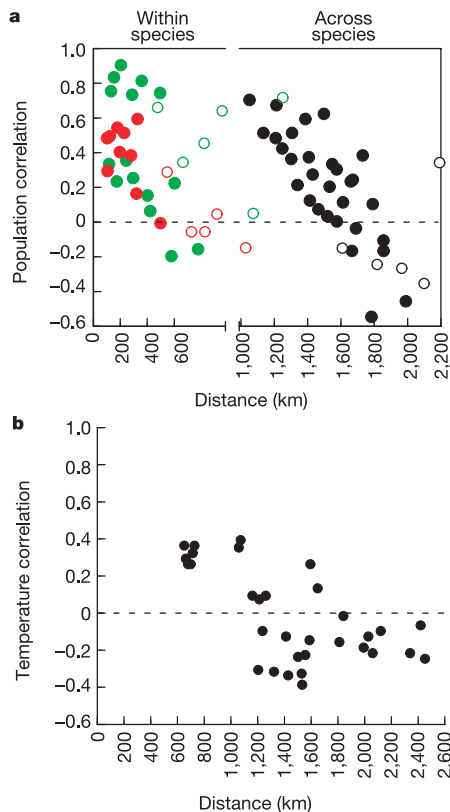


Figure 2 Distance-dependent population synchrony and weather correlations. **a**, The spatial dimension of within-species synchrony between pairs of populations of caribou (green dots; $r = -0.59$, $P = 0.01$) and musk oxen (red dots; $r = -0.60$, $P = 0.03$), and of cross-species synchrony between caribou and musk oxen populations on opposite coasts of Greenland (black dots; $r = -0.79$, $P < 0.001$). Shown are all possible lag-0 Pearson correlation coefficients between caribou and musk oxen populations located in Fig. 1a plotted against the euclidean distance (Methods) between populations. **b**, Distance-dependent correlation in local winter temperatures on opposite coasts of Greenland. Shown are all possible pairs of lag-0 Pearson correlation coefficients between weather stations versus the euclidean distance between stations ($r = -0.65$, $P < 0.001$). Horizontal dotted line indicates zero correlation; open circles represent the correlations including the two populations that are not influenced by the NAO (Methods); inclusion of these populations resulted in the following correlations: caribou, $r = -0.21$, $P > 0.20$; musk oxen, $r = -0.85$, $P < 0.001$; cross-species, $r = -0.73$, $P < 0.001$.

pairwise weather correlation is not meaningful. Therefore, we plotted, for each pair of cross-species populations, their correlation coefficient against the ratio of the coefficients quantifying the influence of the NAO on the dynamics of each population¹⁸.

The NAO effect ratio explains 42% of the variation in the degree of cross-species correlation, and a scatter plot of this relationship indicates that the more similar the strength of the effect of the NAO on any pair of populations, the more highly correlated are their dynamics (Fig. 3). Notably, as the magnitude of the effect of the NAO on any pair of populations diverges, so does the magnitude of their correlation with each other (Fig. 3). In the extreme, populations influenced by the NAO in opposing directions (resulting in a negative NAO effect ratio) are increasingly out of phase with each other (Fig. 3).

Because the influence of the NAO on the dynamics of populations in each species varies with latitude¹⁸, we also investigated the influence of latitudinal differences between pairs of caribou–musk oxen populations on cross-species synchrony in a multivariate regression model. This analysis explained an additional 32% of the variation in cross-species population correlation (total $R^2 = 0.74$, $P < 0.001$) and indicated that populations at proximate latitudes are more highly correlated than those at distant latitudes. We propose that this effect may reflect the influences of latitudinal gradients in plant phenology or forage productivity on population dynamics, but currently lack data to test these hypotheses.

Spatial coupling of widely distributed populations by large-scale climate systems as documented here indicates that climate change can influence, and already might be influencing, the dynamics of several populations and several species simultaneously. Indeed, it is difficult to explain the correlated dynamics of caribou and musk oxen that are separated by a continent-wide ice cap by means other than climate. Likewise, the concordance of the correlations between synchrony in each species and the NAO index (Fig. 1b) with the correlations between local winter temperatures and the NAO index within the range of each species (Fig. 1a) seems to implicate the NAO as the large-scale climate system that underlies population synchrony in this case. Although it may seem that the results of our analyses are particular to the climatic conditions of Greenland, the NAO does not appear to explain more of the variance in interannual density fluctuations of these populations than those of other, temperate populations of large mammals^{12–14,18}. Further analyses will help discern the ubiquity of the role of the NAO, or other large-scale climate systems, in cross-species synchrony. □

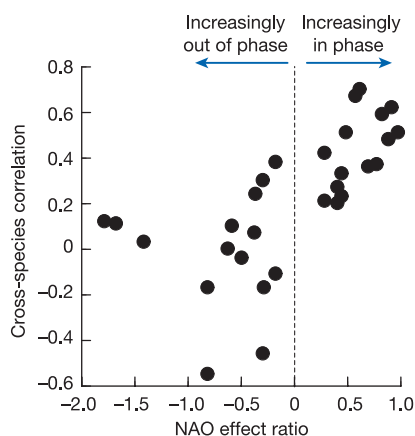


Figure 3 Relation between the degree of pairwise cross-correlation between caribou and musk oxen populations and the ratio of the effect of the NAO on the dynamics of each population ($r = 0.65$, $P < 0.001$). x-axis data were calculated from the coefficients of the terms quantifying the strength of the influence of the NAO on the dynamics of each population, according to the most parsimonious autoregressive models¹⁸.

Methods

Population data

Population indices were derived for the two focal species from harvest data compiled by the Greenlandic Ministry of Denmark for caribou (1908–1981) throughout west Greenland²⁵, and survey data on local musk oxen densities (1960–1988) throughout northeast Greenland compiled by the Danish Sirius patrol¹⁹. The caribou data derive from seven geographically isolated locales separated by fjords and ice fields along the whole west coast of Greenland, whereas the musk oxen data derive from six similarly isolated locales along the northeast coast of Greenland (Fig. 1a). The spatio-temporal extent of these data is thus comparable to that of the Hudson Bay fur trading data on Canada lynx, mink and muskrat, which have been important in theoretical and statistical modelling of spatial synchrony^{8,16,26,27}. We assume that the harvest data provide an index of caribou populations that scales roughly linearly with population size but, in the absence of independent data on actual censuses, we cannot be certain that this is the case.

NAO influence on populations

As Moran specified¹, if two populations showing similar intrinsic dynamics are both influenced by climate, then the degree of correlation between them should scale with the correlation in weather variables influencing their respective dynamics. Hence, in the interest of investigating the contribution of the NAO climate signal to spatial synchrony, the focus of our analysis is the 11 populations (6 caribou and 5 musk oxen) in which density fluctuations are influenced by the NAO. An influence of the NAO was identified in five of six caribou populations¹⁸; we analysed the dynamics of the remaining caribou population (Qeqertarsuaq) using the approach used previously¹⁸ and found a one-year delayed effect of the NAO. We present results of analyses of all 13 populations for comparison.

Population synchrony

To assess spatial patterns of synchrony in the caribou and musk oxen data, as well as cross-species synchrony, we calculated all possible pairs of lag-0 cross-correlation among populations of caribou, musk oxen and the two combined. These were then plotted against the euclidean distance between pairs of populations, estimated using a web-based surface distance calculator (<http://www.vsv.slu.se/johnb/java/lat-long.htm>). Latitude–longitude coordinates were taken at the centre of the geographical distribution of each population. We used weather station data on annual (1967–1994) values of mean winter temperatures, available from nine stations at or near the focal populations, to assess similar spatial patterns in local weather. These data show strong correlation with the NAO index (see ref. 18).

To analyse temporal fluctuations in spatial synchrony in relation to temporal fluctuations in climate, we developed a time series on synchrony among populations in each species. We used the inverse of the coefficient of variation of the log_e (ln)-scale data, calculated each year, as an index of synchrony²⁸. This was regressed against the NAO winter index (<http://www.cgd.ucar.edu/~jhurrell/nao.html>) in linear and quadratic models (the latter to test for possible nonlinear relations between the NAO and synchrony). To avoid spurious correlations owing to temporal trends in the caribou time series²⁵, we included ‘year’ as an independent variable in the analysis of synchrony in caribou. No temporal trend is evident in the musk oxen data¹⁹. For caribou adding a quadratic term did not improve the fit over the linear model ($R^2_{\text{linear}} = 0.43$; $R^2_{\text{quadratic}} = 0.43$), whereas for musk oxen adding a quadratic term improved the fit of the model ($R^2_{\text{linear}} = 0.35$; $R^2_{\text{quadratic}} = 0.45$), which implies that the underlying relation between the NAO and spatial synchrony in musk oxen may be nonlinear. Inclusion of the two populations that are not influenced by the NAO resulted in a poorer model for both caribou ($R^2 = 0.30$) and musk oxen ($R^2 = 0.07$).

Received 21 May; accepted 31 July 2002; doi:10.1038/nature01064.

1. Moran, P. A. P. The statistical analysis of the Canadian lynx cycle. II. Synchronization and meteorology. *Aust. J. Zool.* **1**, 291–298 (1953).
2. Ottersen, G. et al. Ecological effects of the North Atlantic Oscillation. *Oecologia* **128**, 1–14 (2001).
3. Walther, G.-R. et al. Ecological responses to recent climate change. *Nature* **416**, 389–395 (2002).
4. Marshall, J. et al. North Atlantic climate variability: phenomena, impacts, and mechanisms. *Int. J. Climatol.* **21**, 1863–1898 (2001).
5. Koenig, W. D. Spatial autocorrelation of ecological phenomena. *Trends Ecol. Evol.* **14**, 22–26 (1999).
6. Ranta, E., Kaitala, V. & Lindström, J. Spatially autocorrelated disturbances and patterns in population synchrony. *Proc. R. Soc. Lond. B* **266**, 1851–1856 (1999).
7. Grenfell, B. T. et al. Noise and determinism in synchronized sheep dynamics. *Nature* **394**, 674–677 (1998).
8. Stenseth, N. C. et al. Common dynamic structure of Canada lynx populations within three climatic regions. *Science* **285**, 1071–1073 (1999).
9. Schwartz, M. K., Mills, L. S., McKelvey, K. S., Ruggiero, L. F. & Allendorf, F. W. DNA reveals high dispersal synchronizing the population dynamics of Canada lynx. *Nature* **415**, 520–522 (2002).
10. Hurrell, J. W. Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. *Science* **269**, 676–679 (1995).
11. Forchhammer, M. C., Post, E. & Stenseth, N. C. Breeding phenology and climate. *Nature* **391**, 29–30 (1998).
12. Forchhammer, M. C., Stenseth, N. C., Post, E. & Langvatn, R. Population dynamics of Norwegian red deer: density-dependence and climatic variation. *Proc. R. Soc. Lond. B* **265**, 341–350 (1998).
13. Post, E. & Stenseth, N. C. Climatic variability, plant phenology, and northern ungulates. *Ecology* **80**, 1322–1339 (1999).
14. Milner, J. M., Elston, D. A. & Albon, S. D. Estimating the contributions of population density and climatic fluctuations to interannual variation in survival of Soay sheep. *J. Anim. Ecol.* **68**, 1235–1247 (1999).
15. Forchhammer, M. C., Clutton-Brock, T. H., Lindström, J. & Albon, S. D. Climate and population

- density induce long-term cohort variation in a northern ungulate. *J. Anim. Ecol.* **70**, 721–729 (2001).
16. Haydon, D. T., Stenseth, N. C., Boyce, M. S. & Greenwood, P. E. Phase coupling and synchrony in the spatiotemporal dynamics of muskrat and mink populations across Canada. *Proc. Natl Acad. Sci. USA* **98**, 13149–13154 (2001).
 17. Born, E. W. & Böcher, J. *The Ecology of Greenland* (Atuakkiortik, Nuuk, 2001).
 18. Forchhammer, M. C., Post, E., Stenseth, N. C. & Boertmann, D. Long-term responses in arctic ungulate dynamics to variation in climate and trophic processes. *Pop. Ecol.* **44**, 113–120 (2002).
 19. Forchhammer, M. & Boertmann, D. The muskoxen, *Ovibos moschatus*, in north and northeast Greenland—population trends and the influence of abiotic parameters on population dynamics. *Ecography* **16**, 299–308 (1993).
 20. Klein, D. R. The roles of climate and insularity in establishment and persistence of *Rangifer tarandus* populations in the high Arctic. *Ecol. Bull.* **47**, 96–104 (1999).
 21. Skogland, T. *Comparative Social Organization of Wild Reindeer in Relation to Food, Mates, and Predator Avoidance* (Paul Parey, Berlin, 1989).
 22. Post, E. & Klein, D. R. Caribou calf production and seasonal range quality during a population decline. *J. Wildl. Manage.* **63**, 335–345 (1999).
 23. Hedges, L. V., Gurevitch, J. & Curtis, P. The meta-analysis of response ratios in experimental ecology. *Ecology* **80**, 1150–1156 (1999).
 24. Blasius, B. & Stone, L. Ecology—nonlinearity and the Moran effect. *Nature* **406**, 846–847 (2000).
 25. Meldgaard, M. The Greenland caribou—zoogeography, taxonomy, and population dynamics. *Meddelelser om Grønland* **20**, 1–88 (1986).
 26. Ranta, E., Kaitala, V. & Lundberg, P. The spatial dimension in population fluctuations. *Science* **278**, 1621–1623 (1997).
 27. Ranta, E., Kaitala, V., Lindström, J. & Helle, E. The Moran effect and synchrony in population dynamics. *Oikos* **78**, 136–142 (1997).
 28. Post, E., Levin, S. A., Iwasa, Y. & Stenseth, N. C. Reproductive asynchrony increases with environmental disturbance. *Evolution* **55**, 830–834 (2001).
 29. Lodwick, G. D. & Whittle, J. A technique for automatic contouring of field survey data. *Aust. Comp. J.* **2**, 104–109 (1970).
 30. SPSS Inc. *SYSTAT for Windows, version 8.0* (SPSS, Chicago, 1998).
 31. Post, E. Large-scale climate synchronizes the timing of flowering by multiple species. *Ecology* (in the press).

Acknowledgements We thank O. Bjørnstad and W. Koenig for analytical advice; and T. Coulson and E. Ranta for comments. E.P. acknowledges the financial support of the National Science Foundation and the Environmental Consortium at Pennsylvania State University. M.C.F. acknowledges the financial support of the Danish National Science Research Council.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for material should be sent to E.P. (e-mail: esp10@psu.edu).

Selective habituation shapes acoustic predator recognition in harbour seals

Volker B. Deecke^{*†}, Peter J. B. Slater^{*} & John K. B. Ford^{†‡}

^{*} School of Biology, University of St Andrews, Scotland, KY16 9TS, UK
[†] Vancouver Aquarium Marine Science Centre, Vancouver, British Columbia, V6B 3X8, Canada
[‡] Pacific Biological Station, Department of Fisheries and Oceans, Nanaimo, British Columbia, V9T 6N7, Canada

Predation is a major force in shaping the behaviour of animals^{1–3}, so that precise identification of predators will confer substantial selective advantages on animals that serve as food to others. Because experience with a predator can be lethal, early researchers studying birds suggested that predator recognition does not require learning^{4,5}. However, a predator image that can be modified by learning and experience will be advantageous in situations where cues associated with the predator are highly variable or change over time. In this study, we investigated the response of harbour seals (*Phoca vitulina*) to the underwater calls of different populations of killer whales (*Orcinus orca*). We found that the seals responded strongly to the calls of mammal-eating killer whales and unfamiliar fish-eating killer whales but not to the familiar calls of the local fish-eating population. This demon-

strates that wild harbour seals are capable of complex acoustic discrimination and that they modify their predator image by selectively habituating to the calls of harmless killer whales. Fear in these animals is therefore focused on local threats by learning and experience.

The northeastern Pacific Ocean is home to two distinct forms of killer whales. Resident killer whales live in large stable groups and feed exclusively on fish. Transient killer whales live in smaller social groups and prey only on marine mammals^{6,7}. The two different forms do not interbreed and rarely interact. Resident killer whales along the west coast of North America fall into three distinct communities with adjacent home ranges (Alaskan residents, northern residents and southern residents). Transients, on the other hand, form a continuous population from northern California to southeastern Alaska^{8–11}.

Resident and transient killer whales show striking differences in their vocal behaviour. Residents frequently emit echolocation clicks and communicative calls whereas transients are usually silent^{12,13}. Resident killer whales have a complex system of vocal dialects: different social groups have repertoires of 7–17 structurally distinct stereotyped call types (see Fig. 1b and c). The degree of sharing of call types between groups within a community varies from no shared call types to complete sharing of the repertoire^{14,15}, and the structure of stereotyped call types slowly changes with time¹⁶. In contrast, all members of the transient population share most of their call types. No call types are shared among members of different resident communities or between residents and transients.

Harbour seals are the most commonly taken prey of transient killer whales in the coastal waters of British Columbia, Canada⁶, and predation from transients is likely to be a significant source of mortality. Because harbour seals have good underwater hearing at the frequencies of killer whale vocal communication¹⁷, and because underwater calls of killer whales can be heard over long distances¹⁸, it would be beneficial for harbour seals to respond to the calls of transients with anti-predator behaviour. When the salmon migrate through these waters, groups of resident killer whales, which pose no predatory threat to seals, will often spend several weeks in a

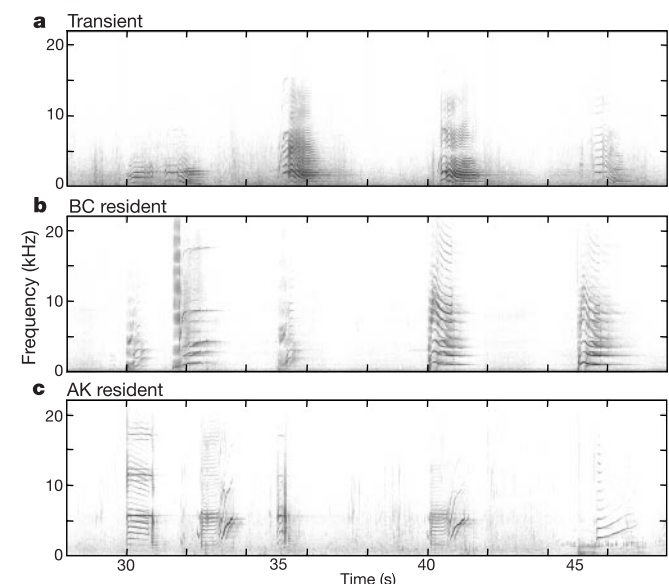


Figure 1 Playback sequences for experiment 2. Spectrograms of sections ($t = 28$ s to $t = 48$ s) of playback sequences illustrating the differences between the calls of mammal-eating killer whales (transients), familiar fish-eating killer whales (British Columbia (BC) residents) and unfamiliar fish-eating killer whales (Alaskan (AK) residents). Several (four transient, seven BC resident and four AK resident) sequences were used for each playback type.

relatively small area¹⁹. Seals would therefore waste time and energy if they responded to all killer whale calls indiscriminately. However, given the complexity of killer whale vocal communication, especially of the resident dialect system, consistent discrimination between calls of the two forms represents a formidable learning task.

We investigated the response of harbour seals to killer whale calls by conducting two playback experiments near reefs where seals haul out onto land off southern British Columbia, Canada. In experiment 1, we used a paired design in which test and control sequences were played in random order at the same seal haulout site on consecutive days, to test whether harbour seals show a significant response to the calls of transient killer whales. The two types of playback sequences only differed in the fact that test sequences contained killer whale calls, whereas control sequences did not. Extraneous stimuli (for example, missed echolocation clicks, or artefacts generated by either the digital editing or the playback equipment) were therefore controlled for (see Methods for details). The number of seals at the surface showed a significantly stronger decrease after the playback of calls of transient killer whales than after control sequences (paired *t*-test, $t_{1,3} = -6.77$, $P < 0.01$). The distance of the nearest seals to the playback source increased after both playback types, and the difference in this measure was not significant (paired *t*-test, $t_{1,3} = 0.83$, $P = 0.47$).

The change in the number of seals shows that harbour seals respond to the calls of transients, either because they perceive them as dangerous or, since transients vocalize only infrequently, because they represent an unfamiliar and hence potentially dangerous cue. Moving away from the surface, where a seal would be very visible and present a good echolocation target, is probably an effective strategy to avoid detection by a visual or echolocating marine predator. This response suggests that transient killer whales pay an ecological cost for vocalizing, because calling decreases their chance of a successful attack and this cost may explain why transient killer whales vocalize less frequently than residents.

In experiment 2, we played calls of transients and British Columbia (BC) residents to see whether seals responded differently to calls of familiar fish-eating and mammal-eating killer whales. To test whether any difference in the response resulted from associative learning (that is, learning to associate the calls of transients with danger) or selective habituation²⁰ (habituating to the calls of residents, because these calls are never followed by an attack), we played calls of Alaskan (AK) residents. These are ecologically similar to BC residents⁷, and occasionally interbreed with northern residents^{10,11}. Because AK residents occur more than 600 km to the north of our study area, seals would be unfamiliar with their calls. If

seals generally do not respond to killer whale calls, but have learned to associate calls of transients with danger, they would not respond to these unfamiliar calls. However, if seals initially responded to all calls, but had habituated to those of the harmless residents, they would respond to unfamiliar calls. Examples of playback sequences are shown in Fig. 1.

The changes in the number of seals visible at the surface (Fig. 2) showed that seals responded differently to the playback of calls of different forms of killer whales (one-way analysis of variance (ANOVA), $F_{2,12} = 12.11$, $P < 0.01$). Seals responded significantly less to calls of familiar fish-eating killer whales than to mammal-eating killer whales (Tukey's test, $P < 0.01$) and unfamiliar fish-eating killer whales (Tukey's test, $P < 0.01$). The responses to mammal-eating and unfamiliar fish-eating killer whales were statistically indistinguishable (Tukey's test, $P = 0.99$). The changes in the distance to the playback source follow the same pattern, but differences are not significant (one-way ANOVA, $F_{2,12} = 0.69$, $P = 0.52$). The different response of harbour seals to the calls of transients and familiar resident killer whales suggests that seals are able to discriminate between the calls of the two populations. Owing to the complexity of killer whale vocal communication, this is not a small feat: since some groups within a resident community have no call types in common¹⁴, the amount of variation among resident groups is as great as the difference between residents and transients. The fact that the seals responded strongly to the calls of fish-eating killer whales with which they had no experience shows that the difference in the response is due to selective habituation to the calls of harmless resident killer whales rather than to associative learning.

Imprecise predator recognition generates fitness costs. However, the currencies of cost for a predator image that is too general and one that is too specific are not the same: an over-general predator image causes an animal to respond to harmless cues and thus waste time or energy. In contrast, an over-specific predator image leads to missed detections of predators, thus posing a risk to life. Animals are limited in their sensory abilities and cannot at the same time minimize the probability of a false alarm while maximizing the probability of correct detection of predators^{21,22}. Because losing a life is a far higher penalty than losing a meal or a mating opportunity, animals should maximize the probability of correct detection of predators at the expense of false alarms.

Having a learned component to predator recognition is advantageous in situations where the nature of the predatory threat is to some degree unpredictable²³, or where cues associated with the predator are highly variable or change with time. Animals that adjust their predator image by associative learning start out with one that is too specific and expand it by adding cues (either when attacked after sensing the cue, or by seeing conspecifics respond to it²³) to reflect the actual predators present. This is risky, as it requires experience with the predators. In contrast, selective habituation predicts that prey animals start out with a rather general predator image from which certain harmless cues are removed by habituation. This initially generates costs from false alarms but not from missed detections. It does not require experience with the predator, since any unusual cue that falls within a certain predator class elicits a response. By selective habituation—learning what not to fear—harbour seals pursue the more conservative, and thus more advantageous, strategy. □

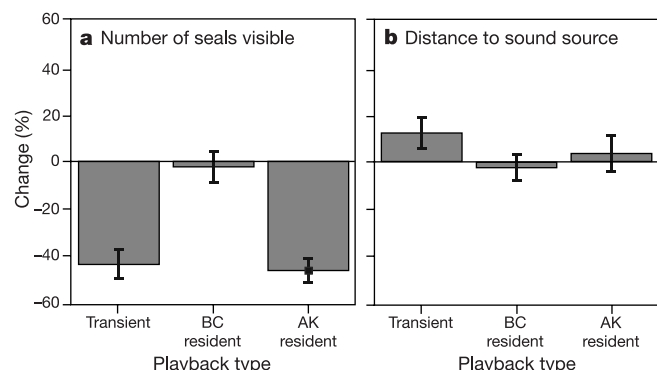


Figure 2 Response of harbour seals to the calls of different killer whale populations. Change in the number of seals visible at the surface (**a**) and the distance of the nearest seal to the playback source (**b**) after playbacks of calls of mammal-eating (Transient), familiar fish-eating (BC resident) and unfamiliar fish-eating (AK resident) killer whales (experiment 2). Ten trials were conducted at different seal haulout sites for each playback type. Error bars give mean $\pm$ 1 s.e.

Methods

General playback procedure

Playbacks were performed to harbour seals in the water using a TCD-D8 DAT-recorder (Sony) and an LL916 underwater speaker (Lubell Labs) deployed at a depth of approximately 5 m from a small boat (6 m long aluminium vessel or 4 m long inflatable) anchored about 100 m from a seal haulout. The maximum source level of the loudest call in sequences containing killer whale calls was 148 dB (reference pressure 1 μ Pa at 1 m). This is 5 dB lower than the average source level of discrete calls of resident killer whales¹⁸. We counted the number of seals in the water and measured the distance to the nearest animal with laser rangefinders at 20 s intervals for at least 2 min preceding the playback of

the calls. We played the 1-min playback sequence a single time and continued to note number of animals and distance to the nearest animal for an additional 2 min or more. Before playbacks the average number of seals at the surface was 13.6 (standard deviation, std, 7.9), and the average distance of the nearest seal to the playback source was 60.2 m (std 21.3 m). Trials where fewer than five seals were present before the playback were excluded from the analysis. The strength of the response was expressed as the percentage change in the average number of seals and average distance to the nearest seal from the 2 min before to the 2 min after the calls were played. Playback experiments were conducted off northern Vancouver Island in Johnstone and Queen Charlotte Straits and off southern Vancouver Island in Haro and Georgia straits.

Experiment 1

Test and control playbacks were conducted once each at the same haulout in random order at the same tidal height on consecutive days. Both types of playback sequences were based on identical sections of background noise from a digitized recording of transient killer whales digitally spliced into a 1-min sequence. Sections containing whistles, echolocation clicks or pulsed calls were not used for this purpose. To avoid startle responses caused by the sudden onset of unfamiliar background noise, the volume was slowly faded in over the first 30 s of the sequence and faded out during the last 10 s. For test sequences, five killer whale calls from the same recording, belonging to at least three different call types, were spliced into the sequence between the fades. For control sequences an additional five sections of background noise were spliced in instead of the calls. The volume of each sequence pair was adjusted so that the loudest call in the test sequence had a source level of 148 dB (reference pressure 1 μ Pa at 1 m). We generated and used four such pairs of sequences from recordings of different transient groups and played each at two different haulouts. In order to avoid pseudoreplication²⁴, we averaged responses obtained at haulouts where the same pair of playback sequences was played, so that the number of playback sequences, not trials, determined degrees of freedom. We used a paired *t*-test to test for significant differences between responses to test and control.

Experiment 2

For this experiment, we generated three types of playback sequence using the methodology explained above. Sequences for playbacks of familiar fish-eater calls contained five calls from BC resident killer whales. We used calls of northern residents for playbacks off northern Vancouver Island, and those of southern residents off southern Vancouver Island. For playbacks of unfamiliar killer whale calls we generated sequences from recordings of Alaskan residents made in Prince William Sound, Alaska. Sequences of transient calls were those used as test sequences in experiment 1. Except for the familiar fish-eating killer whales, we generated four sequences for each playback type from recordings of different social groups. For familiar fish-eating killer whales, we generated a total of seven playback sequences (three of northern residents and four of southern residents). Ten trials were conducted for each playback type and again, to avoid pseudoreplication, all responses obtained with the same playback sequence were averaged. We used a one-way ANOVA to test for statistical differences between the playback types and used Tukey's honestly significant difference test²⁵ to determine which playback types differed.

Received 5 March; accepted 15 July 2002; doi:10.1038/nature01030.

1. Tuttle, M. D. & Ryan, M. J. Bat predation and the evolution of frog vocalizations in the neotropics. *Science* **214**, 677–678 (1981).
2. Gliwicz, M. Z. Predation and the evolution of vertical migration in zooplankton. *Nature* **320**, 746–748 (1986).
3. Rattenborg, N. C., Lima, S. L. & Amlaner, C. J. Half-awake to the risk of predation. *Nature* **397**, 397–398 (1999).
4. Lorenz, K. Vergleichende Verhaltensforschung. *Verhand. Deutsch. zool. Gesellschaft* **1939**, 69–102 (1939).
5. Tinbergen, N. Social releasers and the experimental method required for their study. *Wilson Bull.* **60**, 6–51 (1948).
6. Ford, J. K. B. et al. Dietary specialization in two sympatric populations of killer whales (*Orcinus orca*) in coastal British Columbia and adjacent waters. *Can. J. Zool.* **76**, 1456–1471 (1998).
7. Saulitis, E., Matkin, C., Barrett-Lennard, L., Heise, K. & Ellis, G. Foraging strategies of sympatric killer whale (*Orcinus orca*) populations in Prince William Sound. *Mar. Mamm. Sci.* **16**, 94–109 (2000).
8. Bigg, M. A., Olesiuk, P. F., Ellis, G. M., Ford, J. K. B. & Balcomb, K. C. Social organization and genealogy of resident killer whales (*Orcinus orca*) in the coastal waters of British Columbia and Washington State. *Rep. Int. Whal. Commission (spec. issue)* **12**, 383–405 (1990).
9. Hoelzel, R. A. & Dover, G. A. Genetic differentiation between sympatric killer whale populations. *Heredity* **66**, 191–196 (1991).
10. Hoelzel, A. R., Dahlheim, M. & Stern, S. J. Low genetic variation among killer whales (*Orcinus orca*) in the eastern North Pacific and genetic differentiation between foraging specialists. *J. Hered.* **89**, 121–128 (1998).
11. Barrett-Lennard, L. G. *Population Structure and Mating Patterns of Killer Whale Populations in the Northeastern Pacific, as Revealed by DNA Analysis*. PhD thesis, Univ. British Columbia (2000).
12. Ford, J. K. B. Acoustic behaviour of resident killer whales (*Orcinus orca*) off Vancouver Island, British Columbia. *Can. J. Zool.* **67**, 727–745 (1989).
13. Barrett-Lennard, L. G., Ford, J. K. B. & Heise, K. A. The mixed blessing of echolocation: Differences in sonar use by fish-eating and mammal-eating killer whales. *Anim. Behav.* **51**, 553–565 (1996).
14. Ford, J. K. B. Vocal traditions among resident killer whales (*Orcinus orca*) in coastal waters of British Columbia, Canada. *Can. J. Zool.* **69**, 1454–1483 (1991).
15. Yurk, H., Barrett-Lennard, L., Ford, J. K. B. & Matkin, C. O. Cultural transmission within maternal lineages: Vocal clans in resident killer whales in Southern Alaska. *Anim. Behav.* **63**, 1103–1119 (2002).
16. Deecke, V. B., Ford, J. K. B. & Spong, P. Dialect change in resident killer whales (*Orcinus orca*): Implications for vocal learning and cultural transmission. *Anim. Behav.* **60**, 619–638 (2000).
17. Mohl, B. Auditory sensitivity of the common seal in air and water. *J. Aud. Res.* **8**, 27–38 (1968).

18. Miller, P. J. O. *Maintaining Contact: Design and Use of Acoustic Signals in Killer Whales*, *Orcinus orca*. PhD thesis, Joint Program in Oceanography/Applied Ocean Science and Engineering, Woods Hole Oceanographic Institution and Massachusetts Institute of Technology (2000).
19. Nichol, L. M. & Shackleton, D. M. Seasonal movements and foraging behaviour of northern resident killer whales (*Orcinus orca*) in relation to the inshore distribution of salmon (*Oncorhynchus* spp.) in British Columbia. *Can. J. Zool.* **74**, 983–991 (1996).
20. Schleidt, W. M. Über die Auslösung der Flucht vor Raubvögeln bei Truthühnern. *Naturwissenschaften* **5**, 141–142 (1961).
21. Wiley, R. H. in *Behavioral Mechanisms in Evolutionary Ecology* (ed. Real, L. A.) 157–189 (Univ. Chicago Press, Chicago, IL, 1994).
22. Bradbury, J. W. & Vehrencamp, S. L. *Principles of Animal Communication* (Sinauer, Sunderland, MA, 1998).
23. Curio, E. Proximate and developmental aspects of antipredator behavior. *Adv. Study Behav.* **22**, 135–238 (1993).
24. Kroodsma, D. E. Suggested experimental design for song playbacks. *Anim. Behav.* **37**, 600–609 (1989).
25. Zar, J. C. *Biostatistical Analysis* (Prentice Hall, Upper Saddle River, NJ, 1996).

Acknowledgements We thank the Vancouver Aquarium Marine Science Centre, the BC Killer Whale Adoption Program, the German Academic Exchange Service. We also thank P. Arcese, L. Barrett-Lennard, C. Brignall, J. Borrowman, M. Borrowman, J. deBoeck, N. Dedeluk, G. Ellis, C. Emmons, M. Enstipp, V. Janik, B. Mackay, D. Mackay, A. Morton, R. North, P.-A. Presi, A. Spong, S. Taylor, F. Ugarte, J. Watson, G. Weingartner, R. Williams and H. Yurk.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to V.B.D. (e-mail: vd2@st-andrews.ac.uk).

Graded persistent activity in entorhinal cortex neurons

Alexei V. Egorov*, Bassam N. Hamam*, Erik Fransén†, Michael E. Hasselmo‡ & Angel A. Alonso*

* Department of Neurology and Neurosurgery, Montreal Neurological Institute and McGill University, Montreal, Quebec H3A 2B4, Canada

† Department of Numerical Analysis and Computer Science, Royal Institute of Technology, S-100 44 Stockholm, Sweden

‡ Department of Psychology, Program in Neuroscience and Center for Memory and Brain, Boston University, Boston, Massachusetts 02215, USA

Working memory represents the ability of the brain to hold externally or internally driven information for relatively short periods of time^{1,2}. Persistent neuronal activity is the elementary process underlying working memory but its cellular basis remains unknown. The most widely accepted hypothesis is that persistent activity is based on synaptic reverberations in recurrent circuits. The entorhinal cortex in the parahippocampal region is crucially involved in the acquisition, consolidation and retrieval of long-term memory traces for which working memory operations are essential². Here we show that individual neurons from layer V of the entorhinal cortex—which link the hippocampus to extensive cortical regions³—respond to consecutive stimuli with graded changes in firing frequency that remain stable after each stimulus presentation. In addition, the sustained levels of firing frequency can be either increased or decreased in an input-specific manner. This firing behaviour displays robustness to distractors; it is linked to cholinergic muscarinic receptor activation, and relies on activity-dependent changes of a Ca²⁺-sensitive cationic current. Such an intrinsic neuronal ability to generate graded persistent activity constitutes an elementary mechanism for working memory.

The entorhinal cortex (EC) is a crucial component of the medial temporal-lobe memory system^{4,5}. EC neurons have been shown to display persistent activity during the delay phase of delayed match or non-match to sample memory trials^{6,7} and the hippo-

campal–parahippocampal region is important in paired-associate learning^{6,8–10}.

Layer V neurons of the EC receive convergent sensory input from cortex^{11,12}, are the target of hippocampal output¹³, and give rise to a massive backward projection to the cortex³. EC layer V is also densely innervated by cholinergic fibres from the basal forebrain¹⁴. Cholinergic muscarinic influences have been shown to be vital to memory processes¹⁵ and muscarinic actions have profound effects on the intrinsic firing pattern of neurons¹⁶; we therefore investigated by means of intracellular recordings in a rat EC slice preparation whether muscarinic actions could induce mnemonic activity in EC layer V cells.

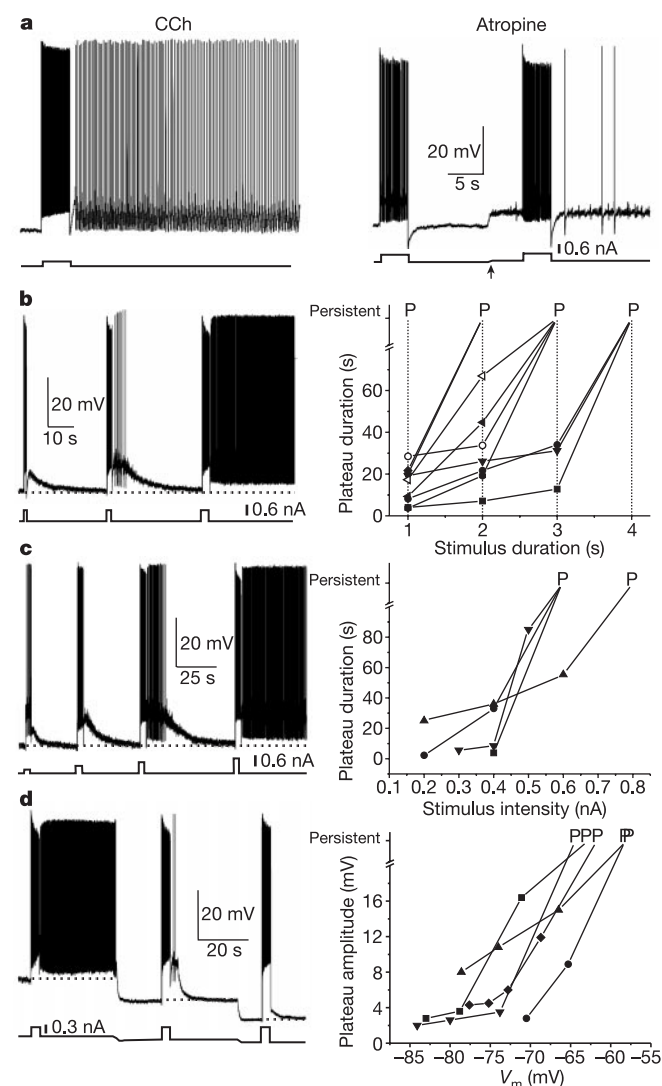


Figure 1 Muscarinic-dependent persistent activity. **a**, CCh-induced persistent firing (5 μ M; left) and block by atropine (1 μ M; right). The arrow below the right current trace signals an imposed direct current (d.c.) shift. **b**, **c**, Left panels show responses to current steps of increasing duration (**b**) and amplitude (**c**) in a representative neuron (CCh, 10 μ M). Right panels show plots of plateau duration versus stimulus duration (**b**) and intensity (**c**) (different symbols correspond to different neurons; in **b**, closed and open symbols correspond to the same neuron studied at a lower and higher stimulus intensity, respectively). **d**, Left panel shows voltage dependence of persistent firing. Right panel shows plot of plateau-potential amplitude as a function of membrane potential (V_m). In all plots, the letter P indicates the transition to persistent firing. In this and Figs 2 and 4, recordings were performed during neurotransmission block as detailed in the Methods. Initial V_m in **a**, **b**, **c** and **d** is -59 mV, -64 mV, -65 mV and -62 mV.

In this study we found that in electrophysiologically identified EC layer V principal cells¹⁷, bath application of the cholinergic agent carbachol (CCh) (5 μ M, $n = 38$; 10 μ M, $n = 49$) blocked the slow afterhyperpolarization that follows a train of action potentials and, in most cases (84% and 98% in 5 μ M and 10 μ M CCh, respectively), triggered the development of a slow depolarizing afterpotential that could give rise to a plateau potential accompanied by spiking (Fig. 1). This plateau potential relied on the activation of muscarinic receptors since its induction was blocked by 1 μ M atropine ($n = 3$; Fig. 1a) or 1 μ M pirenzepine ($n = 4$) and it could also be induced by bath superfusion with muscarine (5, 10 μ M, $n = 11$). The muscarinic-dependent plateau potential and all of its properties were not caused by local circuit reverberation mechanisms, because we commonly studied them during glutamatergic and GABA-mediated neurotransmission block (as specified in Methods). Nevertheless, the plateau activity could be induced equally well with synaptic stimulation during intact neurotransmission ($n = 7$; see below).

Muscarinic-dependent plateau potentials have been described in several cortical neuronal populations^{18–20} including principal cells from EC layer II²¹, where they could provide a cellular mechanism for the delayed activity observed during working memory tasks²². However, the plateau potentials observed in EC layer II neurons always self-terminate²¹. In contrast, we found that when EC layer V cells were activated from a resting level of about 10 mV or less from spike threshold, increases in the stimulus duration (Fig. 1b) or intensity (Fig. 1c) always led to an increase in the duration of the plateau potential (Fig. 1b and c, right-hand plots), which could then give rise to a stable state of sustained spiking for an apparently indefinite period of time (Fig. 1b–d) (tested up to 13 min, although we typically interrupted it after about 40–60 s). This therefore represents the muscarinic- and activity-dependent induction of a self-sustained state of stable firing. We tested responses to current-step durations that ranged from about 0.3 to 8 s and of intensities that commanded spike trains of 15–40 Hz, which is in the beta/low-gamma range of frequencies characteristic of limbic cortices during active states²³. While there was some cell-to-cell variability with respect to the stimulus parameters that elicited persistent firing (Fig. 1b–d), we found that the phenomenon was very robust as it could be elicited in the vast majority of neurons that expressed plateau potentials (see below) and, in every single cell that expressed persistent firing, it could be re-elicited for as long as the recording was maintained. The plateau potential that sustained persistent firing displayed very pronounced voltage dependence. When stimuli of equivalent strength were presented from increasingly negative resting levels, plateau-potential amplitude decreased sharply with membrane hyperpolarization, and persistent firing could never be elicited by stimulation from voltage levels below about -70 mV (Fig. 1d). The ability of neurons to express persistent firing increased with CCh concentration. At a CCh concentration of 10 μ M, a 4-s-long spike train at 15–40 Hz that was evoked from a resting level of about 10 mV or less from spike threshold almost invariably elicited persistent activity (95% of neuron tested; $n = 39$); we therefore chose 4 s as our standard stimulation duration for most of our analysis. However, stimulus durations in the range of 0.3 to 1 s could also be effective in triggering persistent firing in many neurons (8 out of 11 tested; Supplementary Fig. A). At a CCh concentration of 5 μ M, a stimulus of about 4 s in duration elicited persistent activity in 86% of neurons tested ($n = 28$).

Persistent activity for working memory can directly encode dimensions of input or output signals if it can maintain stable analogue values of activity²⁴. We therefore tested whether repetitive application of a given activating stimulus could give rise to a series of graded increases of stable discharge rates. Thus, might muscarinic actions implement in EC layer V neurons the ability to behave as 'neural integrators'? Indeed, as in Fig. 2a (see also Supplementary Information), repetitive application of an input that would give rise

to a state of sustained firing always led to well-defined increases of stable discharge rates in all neurons tested ($n = 13$ in $10 \mu\text{M}$ CCh and $n = 8$ in $5 \mu\text{M}$ CCh). These increases consisted of three to seven levels up to a ceiling (Fig. 2a) where no further enhancement in firing rate was observed. The average maximum persistent firing frequency induced in this manner was $9.8 \pm 4.6 \text{ Hz}$ ($n = 9$) in $10 \mu\text{M}$ CCh and $8.5 \pm 2.9 \text{ Hz}$ ($n = 7$) in $5 \mu\text{M}$ CCh.

Once persistent firing was initiated we noticed that it could only be turned off by prolonged membrane hyperpolarizations (Figs 2c and 3a). The larger the amplitude of the hyperpolarization, the shorter the time required to turn off the persistent active state, but durations of at least 5–10 s for hyperpolarizations to about 80 mV were required to fully stop persistent firing. Graded increases in firing frequency can be effected by repetitive activating stimuli, so

we examined whether repetitive application of hyperpolarizing current pulse steps would have the opposite effect; that is, lead to graded stable decreases in firing rate ($n = 8$). Indeed, as in the case illustrated in Fig. 2c, discrete decreases in firing rate could always be obtained by repetitive step hyperpolarizations of duration equal to, or longer than, those used to induce persistent firing.

Given the unique phenomenon of intrinsic persistent firing elicited by current-step-driven spike trains, we tested whether local synaptic activation could also lead to a state of persistent firing. As illustrated in Fig. 3a, during intact neurotransmission, plateau potentials that sustained stable firing could either be induced by transiently activating the cells synaptically at about 10–20 Hz or by step depolarizations in all neurons examined ($n = 7$). In addition, graded increases in persistent firing frequency

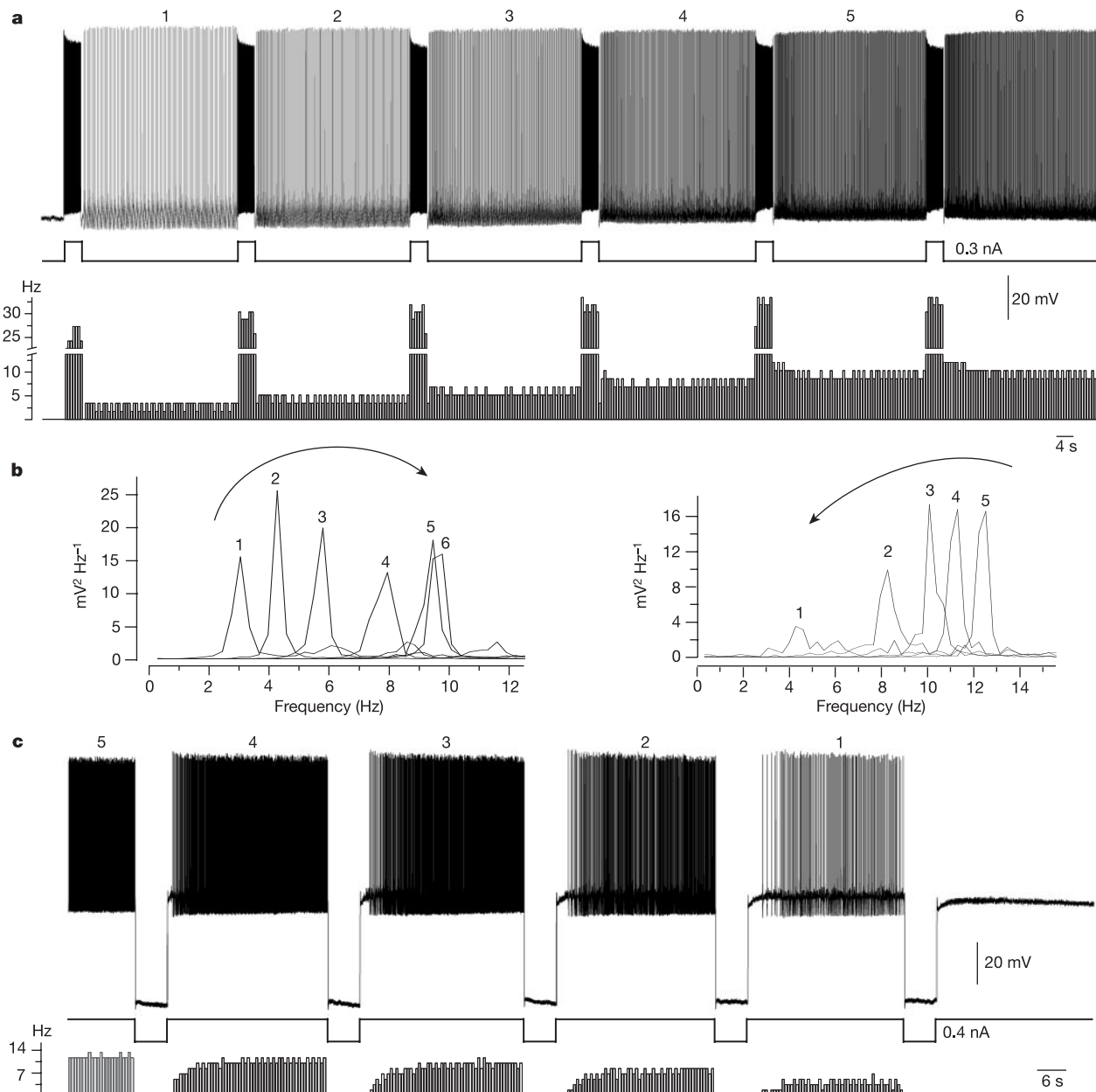


Figure 2 Graded persistent activity. **a**, Repetitive stimulation with a 4-s depolarizing step gives rise to five distinct increases (traces 1 to 6) of stable discharge rate (CCh, $10 \mu\text{M}$). **b**, Fourier analysis plots for the corresponding numbered segments in **a** (left) and **c** (right). **c**, Repetitive application of 6-s hyperpolarizing steps gives rise to discrete decreases of

stable discharge rate and to the eventual cessation of firing (CCh, $5 \mu\text{M}$). The lower diagrams in **a** and **c** correspond to the peristimulus histograms (bin width of 580 ms). Initial V_m in **a** is -64 mV . Final V_m in **b** is -55 mV .

could also be produced by repetitive activation with a synaptic train in all cases tested ($n = 4$) (Fig. 3b). In order to examine whether graded decreases in stable frequency could also be produced by synaptic inhibition, we tested neurons during partial glutamatergic neurotransmission block with 1 mM kynurenic acid. In most cases ($n = 12$ out of 14), we observed that synaptic inhibition was also capable of producing stable decreases in firing frequency (Fig. 3c).

The above data indicate that muscarinic modulation of EC layer V neurons implements in these neurons the internal ability to generate truly persistent activity that can maintain multiple levels of stable firing rate. Another important property of mnemonic persistent activity is that it should be resistant to distracting inputs. We also found that states of stable firing frequency were not affected by relatively brief excitatory-inhibitory stimuli (Supplementary

Fig. B). Typically, current-step-driven spike trains of insufficient strength to elicit persistent firing were also unable to produce graded changes in firing frequency. Similarly, step membrane hyperpolarizations of about 20 mV and shorter than 2 s were always ineffective in causing graded decreases in firing frequency.

Finally, we pharmacologically explored the ionic mechanism underlying the generation of intrinsic persistent activity in layer V cells. The activity-dependent characteristics of the plateau potentials in these neurons clearly suggest that Ca^{2+} influx associated with spiking is an important element. We found that abolishing Ca^{2+} influx by removal of extracellular Ca^{2+} completely and reversibly blocked the muscarinic induced plateau potentials ($n = 4$; Fig. 4a). Similarly, intracellular injection of the Ca^{2+} chelator EGTA also abolished the plateau activity ($n = 3$), thus

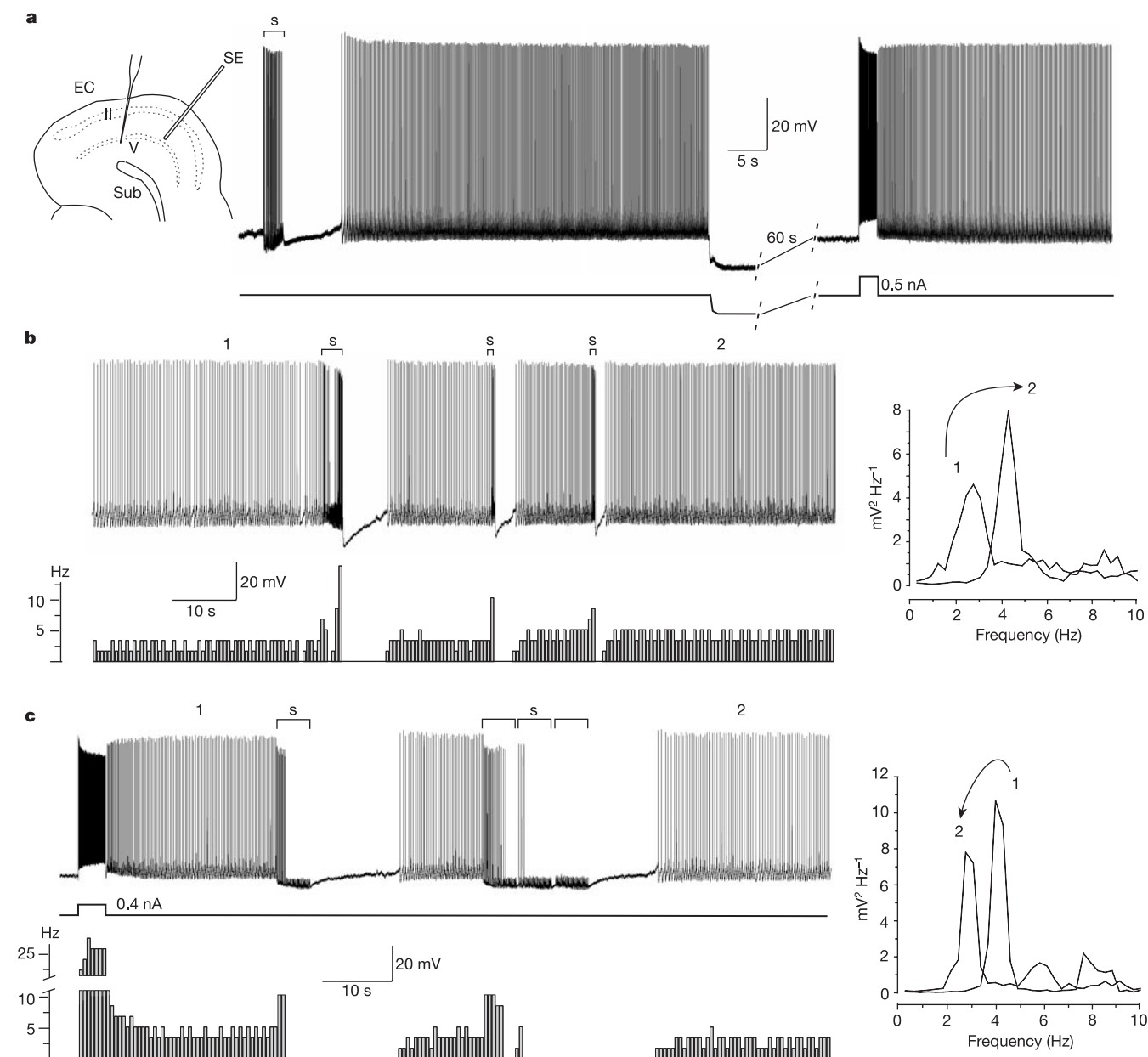


Figure 3 Synaptic induction of persistent activity. **a**, A synaptic train (10 Hz; 2.4 s) induces persistent firing which is then stopped by hyperpolarization and re-initiated by current step depolarization (CCh, 10 μM). Left, schematic representation of the position of recording and stimulating (SE) electrodes. **b**, Stable increase in persistent firing frequency by

synaptic excitation (20 Hz; CCh, 10 μM). **c**, Stable decrease in persistent firing frequency by synaptic inhibition (10 Hz; CCh, 5 μM ; kynurenic acid, 1 mM). Right plots: Fourier analysis for the corresponding labelled segments. Segments labelled as 's' correspond to periods of synaptic stimulation. Initial V_m in **a** and **c** is -62.5 mV and -62 mV.

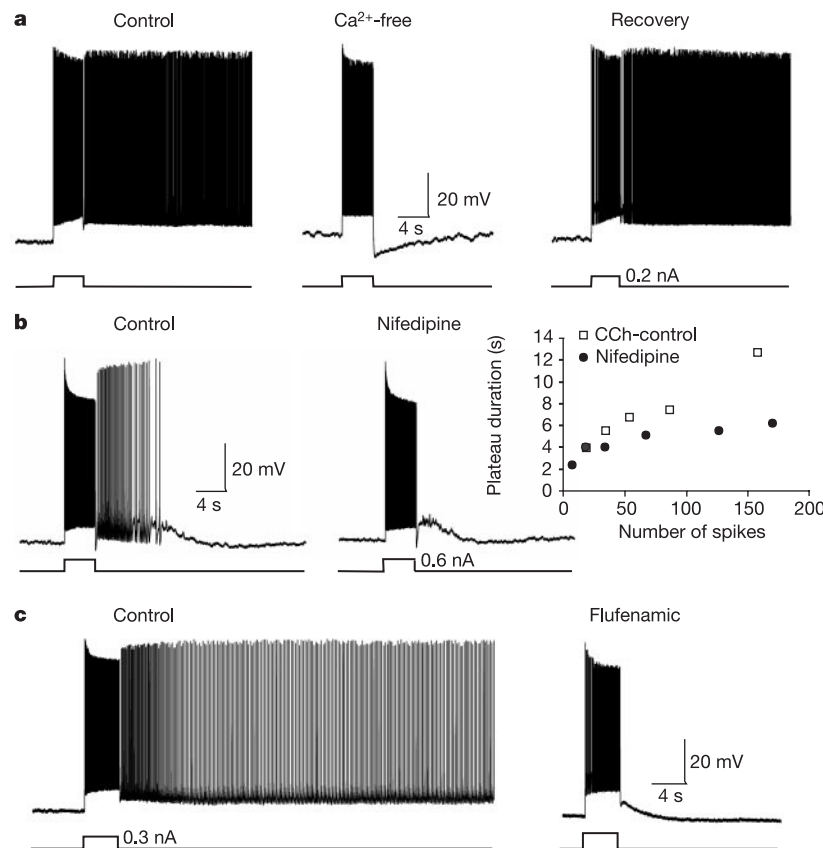


Figure 4 Persistent activity requires activity-dependent Ca²⁺ influx and a non-specific cation current. **a**, Persistent response (left) abolished by removal of extracellular Ca²⁺ (middle) with recovery (right) after Ca²⁺ reintroduction (CCh, 5 μ M). **b**, Persistent activity and plateau potentials were partially blocked by the L-type Ca²⁺ channel blocker

nifedipine (50 μ M; CCh, 5 μ M). The right diagram illustrates the decrease in the slope of the plot of plateau duration versus stimulus intensity (number of spikes in the triggering train) induced by nifedipine. **c**, Complete block of persistent activity by flufenamic acid (10 μ M; CCh, 10 μ M). Initial V_m in **a**, **b** and **c** is -65 mV, -60 mV and -66 mV.

indicating its dependence on intracellular Ca²⁺ rises. Because spiking is expected to lead to Ca²⁺ influx through, at least, high-voltage activated Ca²⁺ channels, we then tested the effect of the L-type Ca²⁺-channel blocker nifedipine (50 μ M) on the plateau activity. In all cases ($n = 6$), nifedipine partially blocked the plateau potentials and curtailed the ability of the cells to generate persistent activity (Fig. 4b). Finally, the possibility that the Ca²⁺-dependent plateau potentials were mediated by a Ca²⁺-activated non-specific cation current was tested: we assessed the effects of the Ca²⁺-activated non-specific current-blocking agent flufenamic acid (10 μ M) (ref. 25), which always ($n = 4$) blocked the cells' ability to generate plateau potentials (Fig. 4c). The above findings point to spike-induced Ca²⁺ influx triggering a slow potential mediated by a cationic current as a basic mechanism for the generation of persistent activity.

Computational modelling studies have shown that recurrent reverberatory circuits can give rise to persistent activity, provided the synaptic feedback has a slow kinetics^{24,26,27}. However, our data *in situ* demonstrate that a non-synaptic spike- and Ca²⁺-induced potential is, in fact, sufficient for the generation of prolonged persistent activity which can be graded in a stimulus-specific manner. EC layer V neurons lie at the core of the hippocampal-neocortical memory system, which implements the acquisition, storage and retrieval of memories for facts and events in a temporally organized and graded manner. The implementation of this type of single-cell mnemonic mechanism, which allows cells to 'hold on' to information for relatively prolonged periods of time and enables the system to perform associational computations,

might be fundamental for the memory operations in the temporal lobe.

Sensory information converges into the EC following a cascade of cortico-cortical projections²⁸. We found that while short stimuli (about 500 ms) could be effective in triggering graded persistent firing, longer stimuli were more effective. This is not surprising since the input to EC layer V neurons may already consist of prolonged discharges from previous stages in the associational hierarchy^{1,2,7}. It has been suggested that while working memory operations in prefrontal cortex may be important for monitoring familiar stimuli, the medial temporal lobe may be more important for matching and active maintenance of new information during memory delays¹⁰. The intrinsic persistent activity displayed by the EC layer V cells represents an ideal mechanism for sustaining information about a new stimulus for memory encoding and/or consolidation purposes. In this respect, EC layer V gives rise to feedback cortical projections, and lesions of the EC have been shown to prevent inferotemporal neurons from representing associations between visual stimuli⁹.

Thus, we have shown that muscarinic cholinergic actions allow EC layer V neurons to behave through a non-synaptic mechanism as analogue memory devices. In contrast to bistable neurons²⁹, they can store multiple bits of information in the form of their activity along a graded dimension determined by stimulus input. We propose that this intrinsic cellular behaviour constitutes an elementary form of mnemonic process on which associative network mechanisms could build to hold externally or internally driven sensory representations. □

Methods

Preparation of brain slices

Brain slices were obtained from male Long Evans rats (150–250 g, Charles River, Canada) using standard procedures²¹. Briefly, animals were deeply anaesthetized with halothane and decapitated; the brain was then rapidly removed, and placed in a cold (4–6 °C) oxygenated Ringer solution containing (in mM): 124 NaCl, 3 KCl, 1.6 CaCl₂, 1.8 MgSO₄, 26 NaHCO₃, 1.25 NaH₂PO₄ and 10 glucose (pH maintained at 7.4 by saturation with 95%O₂/5%CO₂). A block of tissue containing the retrohippocampal region was then dissected out and 400-µm-thick horizontal slices of the entorhinal cortex (EC) were obtained using a vibratome (Pelco). Slices were then incubated in an interface holding chamber at room temperature for >1 h before use. Individual slices were transferred to the interface recording chamber (Fine Scientific tools) one by one, superfused with Ringer solution at a rate of 1–2 ml min⁻¹, and maintained at 34 °C ± 1 °C. Layer V of the entorhinal cortex was identified with a dissecting microscope by transillumination.

Recording procedures, drugs and analysis

Intracellular recordings were obtained using sharp microelectrodes pulled on a Brown-Flaming puller P-87 (Sutter Instruments) from 1.5 mm borosilicate glass (Sutter Instruments). Electrodes were backfilled with 2 M K⁺-acetate (tip resistance of 90–120 MΩ). Signals were amplified using an Axoclamp 2A amplifier, digitized at 5 KHz (Digidata 1200, Axon Inst.) and stored on a Pentium computer, and visualized using AxoScope software (Axon Inst.). Most recordings were performed during glutamatergic and GABA-mediated neurotransmission block with drug cocktails consisting of either a mixture of kynurenic acid (1–10 mM) and picrotoxin (100 µM) (*n* = 81), or a mixture of DL-2-amino-5-phosphonovaleric acid (50 µM), 6-cyano-7-nitroquinoxaline-2,3-dione (10 µM) and picrotoxin (100 µM) (*n* = 7). Synaptic transmission was blocked in most experiments to ensure that the phenomena studied were independent of synaptic transmission. Carbachol, muscarine, atropine and pirenzepine were bath-applied at the desired concentrations from stock solutions (10 mM for CCh and 10 mM for muscarine) in distilled water. Because the muscarinic phenomena studied did not desensitize, in many cases the neurons were directly impaled in the presence of CCh. The resting membrane potential of the neurons was -70.5 ± 3 mV (*n* = 17) in control Ringer, and -63.7 ± 5 mV (*n* = 36) and -62.8 ± 5.5 mV (*n* = 47) in the presence of 5 and 10 µM CCh, respectively. We chose these levels of CCh because in a previous study using the same EC slice preparation we found that higher CCh concentrations led to the production of epileptiform events³⁰. Flufenamic acid and nifedipine were applied from stock solutions made in DMSO so that the final concentration of DMSO in Ringer was ≤0.1%. Control experiments revealed no measurable effects of DMSO on cellular properties or cholinergic modulatory actions (*n* = 4). The Ca²⁺-free solution contained 6 mM Mg²⁺ and 1 mM ethylene glycol-bis-(β-aminoethyl ether) N,N,N',N'-tetraacetic acid (EGTA). For intracellular Ca²⁺ chelation, EGTA was included in the recording micropipette at a concentration of 200 mM. All drugs were purchased from Sigma. Extracellular concentric bipolar electrodes (FHC) were used to induce synaptic responses by local stimulation. Electrophysiological data were analysed using Clampfit 8.2 (Axon Inst.), Origin 6.0 (Microcal) and Matlab (Mathworks) software packages. Spectral (Fourier) analysis was conducted using Origin (spectral resolution 0.305 Hz) and the peristimulus histogram was plotted with Matlab software.

Received 8 August; accepted 20 September 2002; doi:10.1038/nature01171.

- Goldman-Rakic, P. S. Cellular basis of working memory. *Neuron* **14**, 477–485 (1995).
- Fuster, J. M. Network memory. *Trends Neurosci.* **20**, 451–459 (1997).
- Insausti, R., Herrero, M. T. & Witter, M. P. Entorhinal cortex of the rat: cytoarchitectonic subdivisions and the origin and distribution of cortical efferents. *Hippocampus* **7**, 146–183 (1997).
- Scoville, W. B. & Milner, R. Loss of recent memory after bilateral hippocampal lesions. *J. Neurol. Neurosurg. Psychiatr.* **20**, 11–21 (1957).
- Squire, L. R. & Zola-Morgan, S. The medial temporal lobe memory system. *Science* **253**, 1380–1386 (1991).
- Young, B., Otto, T., Fox, G. D. & Eichenbaum, H. Memory representation within the parahippocampal region. *J. Neurosci.* **17**, 5183–5195 (1997).
- Suzuki, W. A., Miller, E. K. & Desimone, R. Object and place memory in the macaque entorhinal cortex. *J. Neurophysiol.* **78**, 1062–1081 (1997).
- Bunsey, M. & Eichenbaum, H. Critical role of the parahippocampal region for paired-associate learning in rats. *Behav. Neurosci.* **107**, 740–747 (1993).
- Higuchi, S. & Miyashita, Y. Formation of mnemonic neuronal responses to visual paired associates in inferotemporal cortex is impaired by perirhinal and entorhinal lesions. *Proc. Natl Acad. Sci. USA* **93**, 739–743 (1996).
- Stern, C. E., Sherman, S. J., Kirchhoff, B. A. & Hasselmo, M. E. Medial temporal and prefrontal contributions to working memory tasks with novel and familiar stimuli. *Hippocampus* **11**, 337–346 (2001).
- Room, P. & Groenewegen, H. J. Connections of the parahippocampal cortex. I. Cortical afferents. *J. Comp. Neurol.* **251**, 415–450 (1986).
- Rempel-Clower, N. L. & Barbas, H. The laminar pattern of connections between prefrontal and anterior temporal cortices in the Rhesus monkey is related to cortical structure and function. *Cereb. Cortex* **10**, 851–865 (2000).
- Naber, P. A., Lopes da Silva, F. H. & Witter, M. P. Reciprocal connections between the entorhinal cortex and hippocampal fields CA1 and the subiculum are in register with the projections from CA1 to the subiculum. *Hippocampus* **11**, 99–104 (2001).
- Alonso, J. R. & Amaral, D. G. Cholinergic innervation of the primate hippocampal formation. I. Distribution of choline acetyltransferase immunoreactivity in the *Macaca fascicularis* and *Macaca mulatta* monkeys. *J. Comp. Neurol.* **355**, 135–170 (1995).
- Hasselmo, M. E. Neuromodulation: acetylcholine and memory consolidation. *Trends Cogn. Sci* **3**, 351–359 (1999).

- Krnjevic, K. Central cholinergic mechanisms and function. *Prog. Brain Res.* **1993**, 285–292 (1993).
- Hamam, B. N., Kennedy, T. E., Alonso, A. & Amaral, D. G. Morphological and electrophysiological characteristics of layer V neurons of the rat medial entorhinal cortex. *J. Comp. Neurol.* **418**, 457–472 (2000).
- Andrade, R. Cell excitation enhances muscarinic cholinergic responses in rat association cortex. *Brain Res.* **548**, 81–93 (1991).
- Fraser, D. D. & MacVicar, B. A. Cholinergic-dependent plateau potential in hippocampal CA1 pyramidal neurons. *J. Neurosci.* **16**, 4113–4128 (1996).
- Haj-Dahmane, S. & Andrade, R. Ionic mechanism of the slow afterdepolarization induced by muscarinic receptor activation in rat prefrontal cortex. *J. Neurophysiol.* **80**, 1197–1210 (1998).
- Klink, R. & Alonso, A. Muscarinic modulation of the oscillatory and repetitive firing properties of entorhinal cortex layer II neurons. *J. Neurophysiol.* **77**, 1813–1828 (1997).
- Fransen, E., Alonso, A. A. & Hasselmo, M. E. Simulations of the role of the muscarinic-activated calcium-sensitive nonspecific cation current INCM in entorhinal neuronal activity during delayed matching tasks. *J. Neurosci.* **22**, 1081–1097 (2002).
- Boeijinga, P. H. & Lopes da Silva, F. H. Modulations of EEG activity in the entorhinal cortex and forebrain olfactory areas during odour sampling. *Brain Res.* **478**, 257–268 (1989).
- Wang, X.-J. Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci.* **24**, 427–488 (2001).
- Partridge, L. D. & Valenzuela, C. F. Block of hippocampal CAN channels by flufenamate. *Brain Res.* **867**, 143–148 (2000).
- Lisman, J. E., Fellous, J. M. & Wang, X. J. A role for NMDA-receptor channels in working memory. *Nature Neurosci.* **1**, 273–275 (1998).
- Seung, H. S., Lee, D. D., Reis, B. Y. & Tank, D. W. Stability of the memory of eye position in a recurrent network of conductance-based model neurons. *Neuron* **26**, 259–271 (2000).
- Insausti, R., Amaral, D. G. & Cowan, W. M. The entorhinal cortex of the monkey: II. Cortical afferents. *J. Comp. Neurol.* **264**, 356–395 (1987).
- Marder, E., Abbott, L. F., Turrigiano, G. G., Liu, Z. & Golowasch, J. Memory from the dynamics of intrinsic membrane currents. *Proc. Natl Acad. Sci. USA* **93**, 13481–13486 (1996).
- Dickson, C. T. & Alonso, A. Muscarinic induction of synchronous population activity in the entorhinal cortex. *J. Neurosci.* **17**, 6729–6744 (1997).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements We thank G. Buzsáki, M. Petrides and W. A. Suzuki for comments on the manuscript. This work was supported by the Canadian Institutes of Health Research and the U.S. National Institutes of Mental Health.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.A.A. (e-mail: angel.alonso@mcgill.ca).

The F-box protein Slimb controls the levels of clock proteins Period and Timeless

Brigitte Grima*, Annie Lamouroux*, Elisabeth Chélot*, Christian Papin*, Bernadette Limbourg-Bouchon† & François Rouyer*

* Institut de Neurobiologie Alfred Fessard (NGI, CNRS UPR 2216) and † Centre de Génétique Moléculaire (CNRS UPR 2167), Centre National de la Recherche Scientifique, av. de la terrasse, 91198 Gif-sur-Yvette, France

The *Drosophila* circadian clock is driven by daily fluctuations of the proteins Period and Timeless, which associate in a complex and negatively regulate the transcription of their own genes^{1,2}. Protein phosphorylation has a central role in this feedback loop, by controlling Per stability in both cytoplasmic and nuclear compartments^{3–6} as well as Per/Tim nuclear transfer^{7,8}. However, the pathways regulating degradation of phosphorylated Per and Tim are unknown. Here we show that the product of the *slimb* (*slmb*) gene⁹—a member of the F-box/WD40 protein family of the ubiquitin ligase SCF complex that targets phosphorylated proteins for degradation^{10–13}—is an essential component of the *Drosophila* circadian clock. *slmb* mutants are behaviourally

arrhythmic, and can be rescued by targeted expression of *Slmb* in the clock neurons. In constant darkness, highly phosphorylated forms of the *Per* and *Tim* proteins are constitutively present in the mutants, indicating that the control of their cyclic degradation is impaired. Because levels of *Per* and *Tim* oscillate in *slmb* mutants maintained in light:dark conditions, light- and clock-controlled degradation of *Per* and *Tim* do not rely on the same mechanisms.

To test whether the SCF-mediated ubiquitin proteasome pathway is involved in the control of *Per* and *Tim* oscillations, we chose to analyse circadian rhythms of flies defective for genes encoding F-box proteins that were known to target phosphorylated substrates for degradation. We started with the *slmb* (*slmb*) gene, which encodes an F-box/WD40 protein regulating transcription factors' levels in the *wingless* and *hedgehog* signalling pathways⁹. *slmb*⁸ mutants that normally die as early larvae¹⁴ were brought to adulthood by providing the *slmb* gene product throughout development under the control of a heat-shock promoter. The rescued *HS-slmb* *slmb*⁸ adult flies, hereafter referred to as *slmb*^m mutants, were then tested for their locomotor activity rhythms in both light:dark (LD) and constant darkness (DD) conditions. *slmb*^m mutants were completely arrhythmic in DD, whereas the heterozygous genotype displayed wild-type rhythms (Table 1). The absence of anatomical defects of the PDF-expressing ventral lateral neurons (LNvs, data not shown), which control the behavioural rhythms^{15,16}, strongly argues against a developmental origin of the mutants' rhythm defect. Furthermore, targeted *slmb* expression using well characterized LNvs-specific *gal4* drivers^{15,16} restored near wild-type activity rhythms, whereas similarly targeted overexpression in a wild-type background lengthened the circadian period (Table 1), indicating a cell-autonomous role of the *slmb* gene in circadian rhythmicity. In LD conditions, *slmb*^m mutants did not display the light-off anticipation of activity that characterizes a functional clock, whereas it was observed in the flies expressing *slmb* under the LNvs-specific *gal1118* driver (Fig. 1). These experiments identify the F-box/WD40

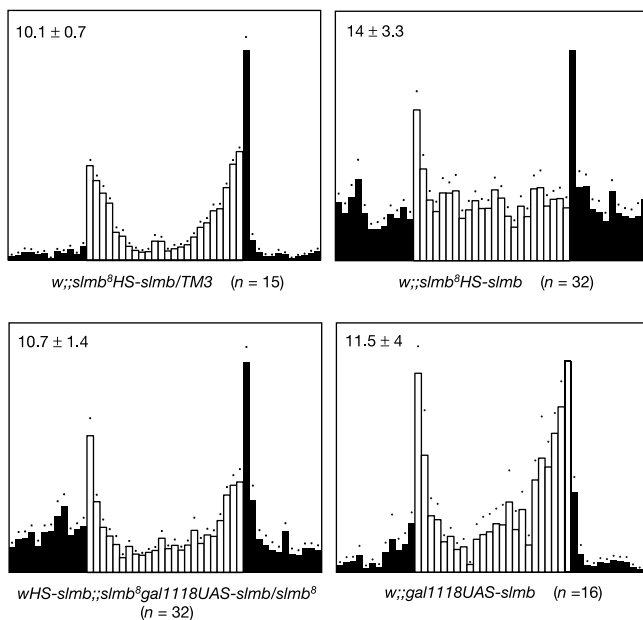


Figure 1 Locomotor activity in LD cycles. Histograms represent the distribution of activity through 24 h, averaged for *n* flies over three days. Open and filled bars indicate day and night phases, respectively. Activity levels are normalized independently for each genotype, and the mean daily activity (number of events per 0.5 h) is given in the upper left corner. Dots indicate the s.e.m. of the activity for each 0.5 h. *w;slmb*⁸ *HS-slmb* and *wHS-slmb;slmb*⁸ are referred to as *slmb*^m in the text and following figures.

protein *Slmb* as an essential component of the *Drosophila* brain clock.

To understand how *Slmb* might affect the circadian oscillator, *slmb*^m mutants were analysed for *Per* and *Tim* oscillations in the head. In wild-type flies maintained in LD cycles, *Per* and *Tim* proteins accumulate and are progressively phosphorylated during night time, with *Tim* disappearing at the end of the night whereas hyper-phosphorylated *Per* persists for a few hours in the morning^{17–19}. A similar temporal pattern persists in DD^{17–19}, and is required to sustain behavioural rhythmicity^{16,20,21}. In contrast, highly phosphorylated *Per* and *Tim* were present at all circadian

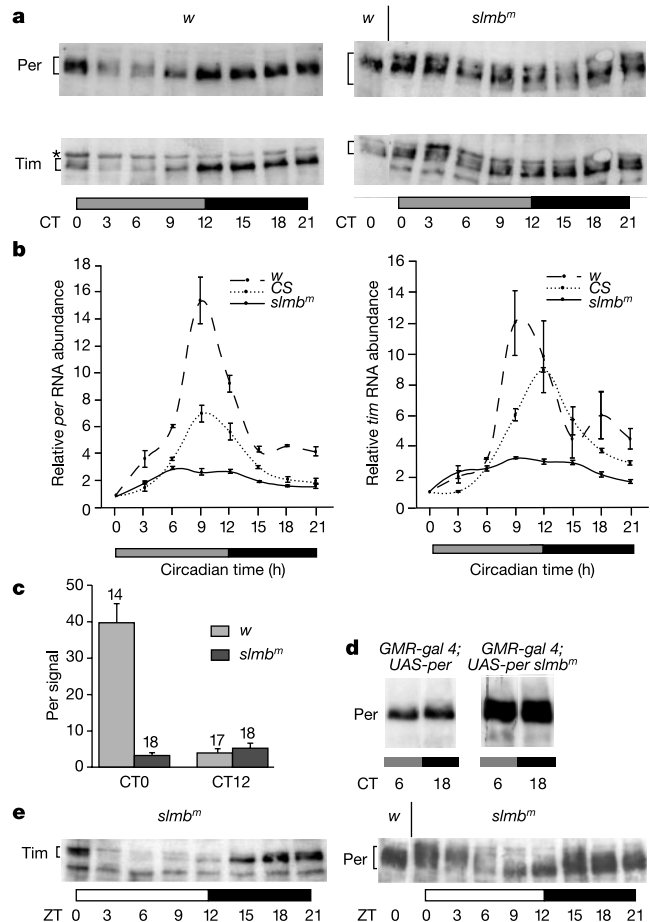


Figure 2 *Per* and *Tim* proteins and RNAs in *slmb*^m mutants. All experiments were done with flies maintained at 20 °C, with no overlap between LD and DD time points. **a**, anti-*Per* and anti-*Tim* western blots from wild-type (*w*) or *slmb*^m head extracts, collected every 3 h in the first day of DD (CT for circadian time, CT 0 is 12 h after the last LD light-off). Asterisk indicates non-specific bands observed in *tim*⁰ flies (not shown). **b**, Relative *per* and *tim* RNA abundance determined by quantitative RT-PCR in *w* and *CS* controls (see Methods) and *slmb*^m head extracts during the first day of DD. Each value is an average of two independent experiments and is given with its associated s.e.m. **c**, *Per* expression in the LNvs of *slmb*^m and *w* controls. Brains were dissected during the first day of DD at CT 0 and CT 12. Compared to the western blots, the low *Per* immunoreactivity in the *slmb*^m dissected brains suggests that the anti-*Per* antibody poorly recognizes the hyper-phosphorylated *Per* that constitutes the main form of the protein in *slmb*^m flies but not in the controls. The quantification of *Per* immunoreactivity is reported in arbitrary units with the associated s.e.m. The number of brain hemispheres analysed is shown above each histogram. **d**, Anti-*Per* western blots from heads of flies overexpressing *Per* in a *slmb*^m or *slmb*⁺ background in DD. **e**, Anti-*Tim* and anti-*Per* western blots from *slmb*^m heads collected every 3 h in LD (ZT for zeitgeber time, ZT 0 is light-on). All the experiments of this figure were done at least twice and gave similar results. Quantifications of the western blots are presented as Supplementary Information.

times in *slmb^m* mutants kept in DD (Fig. 2a, and Supplementary Information), although low-amplitude oscillations of the hypo-phosphorylated forms indicated a weak residual activity of the molecular clock (Fig. 2a). In agreement with the persistence of weak protein cycling in *slmb^m* heads, levels of *per* and *tim* transcripts displayed low-amplitude oscillations (Fig. 2b). We then looked at Per immunoreactivity in the LNVs that control behavioural rhythms (Fig. 2c). At circadian time (CT) 0 and CT 12, which correspond to the peak and trough of Per labelling in *w* flies at 20 °C (Fig. 3e), *slmb^m* mutants showed low levels of Per immunoreactivity, indicating that the oscillations of the proteins levels are also abolished in the clock cells. To determine whether Slmb acts at the protein level or through a transcriptional control, *per* was constitutively over-expressed through a transgene. High-molecular-mass Per proteins were observed to accumulate in head extracts of *slmb^m* but not of

wild-type flies carrying *GMR-gal4* and *UAS-per* transgenes that drive strong Per expression in the eye (Fig. 2d). Altogether, these data indicate that Slmb is involved in the control of phosphorylated Per levels.

In LD conditions, Per and Tim degradation in the morning is driven by both the circadian cycle and by light^{18,19,22,23}. Light-induced Tim degradation involves ubiquitinylation of the protein, and is blocked by proteasome inhibitors²⁴. To test whether Slmb is involved in the light-induced degradation pathway of the clock proteins, Per and Tim levels were assayed in *slmb^m* flies kept in LD conditions. In contrast to constant darkness, robust oscillations of Per and Tim amounts were observed in LD, with both proteins accumulating during the night and showing a strong day-time decrease (Fig. 2e). This shows that light-induced Per and Tim degradation does not occur through the same *slmb*-dependent mechanism as their circadian-cycle-controlled degradation in constant darkness. In addition, the absence of light-off anticipation in the *slmb^m* activity profiles (Fig. 1) suggests that the mutants' altered temporal regulation of phosphorylated Per and Tim does not allow rhythmic outputs to be driven, although protein levels clearly cycle.

Clock-dependent Per and Tim degradation occurs at the end of the circadian cycle, and relieves the transcriptional repression that the proteins exert on their own genes^{1,2}. Per degradation has also been proposed to take place during the rising phase of the protein levels in the early night, and to be responsible for the shift (of 5 h) between *per* messenger RNA and Per protein peaks^{3,4}. In order to determine whether Slmb levels vary during a circadian cycle and may therefore affect Per and Tim only during a limited time window, anti-Slmb antibodies were raised and the Slmb protein was followed in head extracts at different circadian times. A strongly reacting protein, as well as a faintly reacting one slightly above, were detected at a relative molecular mass of 45,000 (M_r 45K) in wild-type flies, and did not show any oscillations of their levels over a 24-h time course (Fig. 3a). Similarly, *slmb* mRNA did not show any cycling (data not shown). Slmb therefore appears not to be circadianly regulated, and could act on different steps of the cycle.

Both early- and late-night Per degradation steps appear to depend upon Per phosphorylation, which requires the casein kinase I encoded by the *double-time* (*dbt*) gene³⁻⁶. To final out how Slmb could affect Per and Tim phosphorylation, we first tested whether Dbt, and Shaggy (Sgg) that has been recently shown to phosphorylate Tim⁷, were affected in *slmb^m* mutants. No alterations of the level or the mobility of these kinases were detected in *slmb^m* head extracts (Fig. 3b). We then investigated whether Slmb could associate with the Per protein, by searching for Per-Slmb interactions in co-immunoprecipitation experiments on head extracts. We found that the Slmb protein was co-precipitated by anti-Per antibodies, and that anti-Slmb could precipitate Per in wild-type flies collected at CT 0 (Fig. 3c). Similar results were obtained with pooled extracts (CT 15-18-21-0-3, not shown). In addition, Slmb co-precipitated with Dbt (Fig. 3c). Because Per, but not Dbt, was profoundly affected in *slmb^m* mutants, our results support Per rather than Dbt as a Slmb target for ubiquitinylation, and suggest that the three proteins constitute a complex. Slmb was co-immunoprecipitated by anti-Per antibodies in *tim⁰* flies (Fig. 3c), indicating that Per-Slmb complexes can form in the absence of Tim. Although twice as much extract was used for *tim⁰* flies to compensate for the low Per levels in this genotype, the amount of immunoprecipitated Slmb suggests that the absence of Tim may favour Per-Slmb complexes. These results fit well with Slmb being involved in the control of unbound Per, either during its cytoplasmic accumulation at the beginning of the protein cycle or during its nuclear degradation at the end⁶. To test whether the formation of Per-Slmb complexes is circadianly controlled, we performed co-immunoprecipitations at the beginning of the night when Per is mostly hypo-phosphorylated, or at the end of the night when Per is highly phosphorylated (Fig. 3d). All time points showed comparable levels of Per-Slmb complexes, and

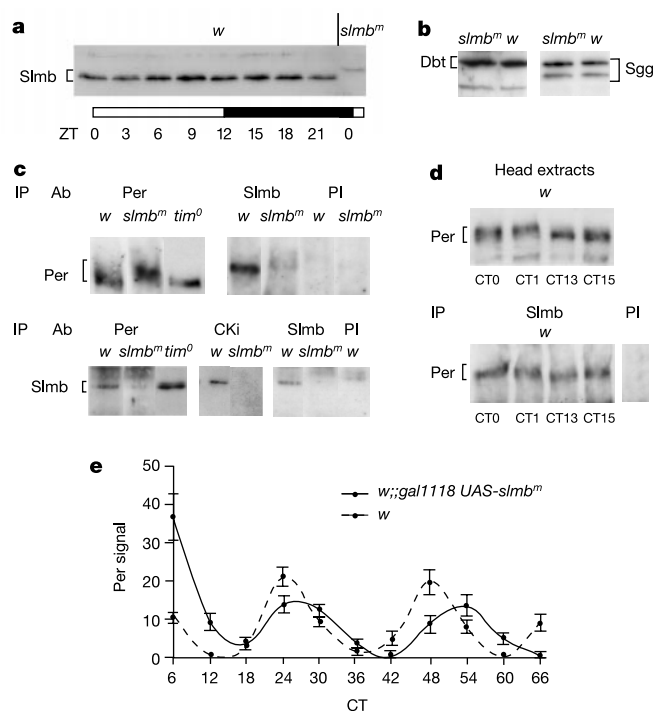


Figure 3 Slmb protein expression and interactions. **a**, Anti-Slmb western blots from wild-type (*w*) head extracts, collected every 3 h in a LD cycle (20 °C). In *slmb^m* head extracts, the wild-type signal disappears but a weak band of higher molecular mass is detected. **b**, Anti-Dbt and anti-Sgg Western blots from *w* and *slmb^m* head extracts. Specific bands are indicated by brackets. **c**, Co-immunoprecipitations from wild-type (*w*), *slmb^m* or *tim⁰* head extracts collected at CT 0. The immunoprecipitation antibody (anti-Per, anti-Slmb, anti-CKI or pre-immune (PI) serum) is indicated above each blot (IP Ab). For immunoprecipitations, anti-Dbt was replaced by anti-human CKI that recognizes the Dbt protein (not shown). Immunoprecipitated proteins were then probed with anti-Per or anti-Slmb antibodies. Blots of total extracts were used as positive controls to define the anticipated IP band (not shown). Adjacent lanes are derived from the same gel (Per western blots of anti-Per and anti-Slmb IP were not migrated similarly, and cannot be compared for Per mobility). A faint Slmb band was detected in anti-Per IP of *slmb^m* extracts, indicating that very low amounts of Slmb are likely to leak from the *HS-slmb* transgene. **d**, Co-immunoprecipitation experiments from wild type at CT 0, 1, 13 and 15. The immunoprecipitation antibody was anti-Slmb, and the precipitates were probed with anti-Per antibody. Western blot from wild-type head extracts indicates the difference of Per phosphorylation between CT 0, 1 and CT 13, 15 at 20 °C. **e**, Per expression in the LNVs of Slmb-overexpressing flies. Flies were entrained for 3 d in LD, and collected in DD from CT6 to CT 66. 6 to 8 brains of each genotype were dissected for each time point. The quantification of Per immunoreactivity is reported in arbitrary units with the associated s.e.m.

Table 1 Locomotor activity rhythms in constant darkness

Genotype	Number of tested flies	Number of rhythmic flies (%)	Mean period of rhythmic flies	Mean power of rhythmic flies
<i>w</i>	40	36 (90)	23.8 ± 0.1	124 ± 6
<i>w;slmb⁸ HS-slmb/TM3</i>	28	28 (100)	24.1 ± 0.1	113 ± 6
<i>w;slmb⁸ HS-slmb</i>	64	7 (11)	29.4 ± 2.9*	42 ± 4
<i>w HS-slmb;slmb⁸</i>	16	1 (6)	28.5†	29†
<i>w HS-slmb;slmb⁸ gal1118 UAS-slmb/slmb⁸</i>	46	43 (93)	25.1 ± 0.1	89 ± 5
<i>w HS-slmb;pdf-gal4/+;slmb⁸ UAS-slmb/slmb⁸</i>	15	14 (93)	25.0 ± 0.3	71 ± 7
<i>w;UAS-slmb</i>	11	10 (91)	24.3 ± 0.1	95 ± 8
<i>w;gal1118</i>	54	51 (94)	25.0 ± 0.1	111 ± 4
<i>w;gal1118 UAS-slmb</i>	70	60 (86)	26.7 ± 0.2	94 ± 6

The mean values of periods and associated powers are given ± s.e.m. *w;slmb⁸ HS-slmb* and *w HS-slmb;slmb⁸* are referred as *slmbtm* in the text.

*The period values of the rhythmic individuals are highly variable for this genotype.

†No s.e.m. as only one rhythmic individual.

several forms of Per were immunoprecipitated by the anti-Slmb antibodies (compare CT 1 and CT 13). This indicates that differently phosphorylated Per molecules can be committed to Per-Slmb complexes.

Possible explanations for the accumulation of highly phosphorylated Per in *slmbtm* mutants would be that partially phosphorylated Per is the relevant Slmb substrate for degradation, or that Slmb targets some Per kinase that is bound to Per. The presence of highly phosphorylated Per in *slmbtm* as early as zeitgeber time (ZT)¹⁵ in LD (Fig. 2e) indicates that Slmb is required for the control of phosphorylated Per accumulation in the early night. Moreover, *Slmb* overexpression in the LNVs resulted in a lengthening of the circadian period (Table 1). In agreement with the behavioural data, *Slmb* overexpression slowed down the oscillations of Per immunoreactivity in these cells, which showed a ~6-h delay compared to wild-type controls after two days (Fig. 3e). These data can be explained by high levels of cytoplasmic Slmb inducing too much degradation of cytoplasmic Per, thus further delaying the night accumulation of the protein, whereas high levels of nuclear Slmb would rather precipitate the fall of the Per protein and shorten the circadian period. We therefore think that Slmb is, at least, involved in the control of cytoplasmic Per accumulation in the early night.

The presence of low-mobility Tim proteins at all circadian times in *slmbtm* mutants (compare with control in Fig. 2a) indicates that the accumulation of phosphorylated Tim is also Slmb-dependent. Remarkably, the Tim kinase Sgg controls the Slmb-dependent proteolysis of Cubitus Interruptus^{9,25,26} and degradation of Armadillo^{9,27}. Our results suggest that phosphorylated Tim could be a Slmb target or that Tim is phosphorylated by a Slmb-dependent kinase. Because Tim is hypo-phosphorylated in *per⁰* flies^{7,19}, it is also possible that the accumulation of hyper-phosphorylated Per in *slmbtm* influences Tim phosphorylation.

Although protein degradation is commonly believed to have a major role in the control of the oscillations of clock proteins, the present work is (to our knowledge) the first to implicate a characterized component of the ubiquitin proteasome pathway. Because cycling of phosphorylated Per proteins also occurs in the mammalian clock^{1,2}, it would be interesting to determine whether the Slmb mammalian homologue β-Trcp¹¹ is involved in the control of phosphorylated Per levels. F-box proteins have been shown to be important at the G1/S transition of the cell cycle, by targeting phosphorylated cyclins and inhibitors of cyclin kinases for degradation by the proteasome¹⁰. Our study therefore suggests that the cell-cycle and the circadian-clock machineries share mechanisms to control the oscillations of phosphorylated proteins. □

Methods

Fly strains

Homozygous *slmb⁸* adults were obtained in the progeny of *slmb⁸ HS-slmb/TM3Sb* flies¹⁴ that were grown at 25°C and heat-shocked daily at 37°C for 1 h, from first-instar larvae until the end of pupariation. Other genotypes homozygous for *slmb⁸* were obtained with the same protocol after the appropriate crossings with the *w;gal1118¹⁶*, *yw;pdf-gal4¹⁵* and

w;GMR-gal4;UAS-per¹⁶ lines. We cannot exclude that residual Slmb is present in the *slmbtm* (*w;slmb⁸ HS-slmb* and *w HS-slmb;slmb⁸*) adults. The *UAS-slmb* construct was made with the *Tslmb* complementary DNA¹⁴ inserted in the pUAST vector and injected into embryos to generate the *w;UAS-slmb* line.

Behavioural analysis

Experiments were carried out with 1–7-d-old males at 20°C in *Drosophila* activity monitors (Trikinetics) as described before¹⁶. For DD analysis, flies were first kept in 12 h:12 h LD cycles over 2–4 d, and recorded in DD over 6 d. Data analysis was done on a Macintosh computer with the Faas software (available upon request), which is derived from the Brandeis Rhythm Package (<http://hawk.bcm.tmc.edu/>). Rhythmic flies were defined by χ^2 periodogram analysis with the following criteria (filter on): power ≥ 20, width ≥ 5, number of peaks ≤ 3.

Western blotting

Head total protein extracts and immunoblottings were made as previously described¹⁶, except that anti-phosphatases (20 mM β-glycerophosphate, 200 μM Na₃VO₄) were added to the extraction buffer. Anti-Slmb antibodies were generated in rats against the C-terminal peptide ENKTGRTPSPALMEH (Neosystem), and used at dilution 1/2,000. The rat anti-Tim antibodies were raised against a GST-Tim (residues 222–557) fusion protein¹⁸ and used at dilution 1/2,000. (GST, glutathione S-transferase) The rabbit anti-Per²⁸ (1/10,000), rat anti-Dbt⁶ (1/500) and Mab2G2C5 anti-Sgg²⁹ antibodies have been already reported.

Immunocytochemistry

Labellings were done on whole-mounted adult brains as previously described¹⁶ with the following modifications: anti-Per serum was applied overnight, Alexa Fluor 594 (Molecular Probes) anti-rabbit IgG goat antibody was used at 1:10,000 as a secondary antibody. Fluorescence intensity was quantified from digital images using the formula: $I = 100 \times (S - B)/B$, that gives the fluorescence percentage above background (S, fluorescence intensity; B, average intensity of the region adjacent to the positive cell).

Immunoprecipitations

Extractions and immunoprecipitations were performed as described⁶, with the following modifications: 2 mg of head total proteins (4 mg for *tim⁰* flies) were extracted in HE buffer supplemented with 20 mM β-glycerophosphate, 200 μM Na₃VO₄, 0.1% gelatin and 1 mM dithiothreitol. 2 μl of immune (or pre-immune) serum and 50 μl of a slurry containing 50% protein G sepharose (Pharmacia) were used for each immunoprecipitation. The UT31 rabbit anti-human CKIe antibody³⁰ was used to immunoprecipitate Dbt. For anti-Slmb immunoprecipitations, rabbit antibodies directed against the same peptide were used instead of rat antibodies. All immunoprecipitations were done at least twice, with similar results.

Quantitative RT-PCR

cDNA were synthesized from 1 μg of head total RNA (Promega SV total RNA isolation system). Quantitative PCR was performed with a Lightcycler using the SYBR green detection protocol (Roche). cDNA samples were mixed with Faststart DNA master SYBR green I mix, 3 mM MgCl₂, 500 nM of each primers in 20 μl and submitted to 40 cycles of PCR (95°C, 15 s; 60°C, 10 sec; 72°C, 20 s). RNA levels were normalized to the levels of *tubulin* RNA and converted to a multiple of the minimal level for each gene (minimal levels were the less variable between experiments). *slmbtm* mutants are in a *w* genetic background but have red eyes owing to the *HS-slmb* transgene. We therefore used both *w* and *Canton S* (CS) flies as controls, because white-eyed flies usually show RNA oscillations of higher amplitude. The primers were 5'TCCTTGTCGCGTGGAACA (exon 1) and 3'CCGAACGAGTGGAGATGAG (exon 2) for *tubulin* (464 bp), 5'CGCCGGACACTGAAAAGG (exon 7) and 3'ATATATAACCTTAGGGCT (exon 8) for *per* (620 bp), and 5'AGTTGGTCATGCGCAGCAAATG (exon 12) and 3'TCCTTTTCGTACACAGATGCCA (exon 13) for *tim* (447 bp).

Received 11 July; accepted 23 August 2002; doi:10.1038/nature01122.

- Allada, R., Emery, P., Takahashi, J. S. & Rosbash, M. Stopping time: the genetics of fly and mouse circadian clocks. *Annu. Rev. Neurosci.* **24**, 1091–1119 (2001).
- Young, M. W. & Kay, S. A. Time zones: a comparative genetics of circadian clocks. *Nature Rev. Genet.* **2**, 702–715 (2001).

3. Kloss, B. *et al.* The *Drosophila* clock gene double-time encodes a protein closely related to human casein kinase 1 ϵ . *Cell* **94**, 97–107 (1998).
4. Price, J. L. *et al.* double-time is a novel *Drosophila* clock gene that regulates PERIOD protein accumulation. *Cell* **94**, 83–95 (1998).
5. Suri, V., Hall, J. C. & Rosbash, M. Two novel doubletime mutants alter circadian properties and eliminate the delay between RNA and protein in *Drosophila*. *J. Neurosci.* **20**, 7547–7555 (2000).
6. Kloss, B., Rothenfluh, A., Young, M. W. & Saez, L. Phosphorylation of period is influenced by cycling physical associations of double-time, period, and timeless in the *Drosophila* clock. *Neuron* **30**, 699–706 (2001).
7. Martinek, S., Inonog, S., Manoukian, A. S. & Young, M. W. A Role for the segment polarity gene shaggy/GSK-3 in the *Drosophila* circadian clock. *Cell* **105**, 769–779 (2001).
8. Curtin, K. D., Huang, Z. J. & Rosbash, M. Temporally regulated nuclear entry of the *Drosophila* period protein contributes to the circadian clock. *Neuron* **14**, 365–372 (1995).
9. Jiang, J. & Struhl, G. Regulation of the Hedgehog and Wingless signalling pathways by the F-box/WD40-repeat protein Slimb. *Nature* **391**, 493–496 (1998).
10. Skowyr, D., Craig, K. L., Tyres, M., Elledge, S. J. & Harper, J. W. F-box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin-ligase complex. *Cell* **91**, 209–219 (1997).
11. Margottin, F. *et al.* A novel human WD protein, h- β TrCP, that interacts with HIV-1 Vpu connects CD4 to the ER degradation pathway through an F-box motif. *Mol. Cell* **1**, 565–574 (1998).
12. Spencer, E., Jiang, J. & Chen, Z. J. Signal-induced ubiquitination of I κ B α by the F-box protein Slimb/ β -TrCP. *Genes Dev.* **13**, 284–294 (1999).
13. Winston, J. T. *et al.* The SCF β -TRCP-ubiquitin ligase complex associates specifically with phosphorylated destruction motifs in I κ B α and β -catenin and stimulates I κ B α ubiquitination in vitro. *Genes Dev.* **13**, 270–283 (1999).
14. Miletich, I. & Limbourg-Bouchon, B. *Drosophila* null slimb clones transiently deregulate Hedgehog-independent transcription of wingless in all limb discs, and induce decapentaplegic transcription linked to imaginal disc regeneration. *Mech. Dev.* **93**, 15–26 (2000).
15. Renn, S. C., Park, J. H., Rosbash, M., Hall, J. C. & Taghert, P. H. A pdf neuropeptide gene mutation and ablation of PDF neurons each cause severe abnormalities of behavioral circadian rhythms in *Drosophila*. *Cell* **99**, 791–802 (1999).
16. Blanchard, E. *et al.* Defining the role of *Drosophila* lateral neurons in the control of circadian activity and eclosion rhythms by targeted genetic ablation and PERIOD protein overexpression. *Eur. J. Neurosci.* **13**, 871–888 (2001).
17. Edery, I., Zwiebel, L. J., Dembinska, M. E. & Rosbash, M. Temporal phosphorylation of the *Drosophila* period protein. *Proc. Natl Acad. Sci. USA* **91**, 2260–2264 (1994).
18. Myers, M. P., Wager-Smith, K., Rothenfluh-Hilfiker, A. & Young, M. W. Light-induced degradation of TIMELESS and entrainment of the *Drosophila* circadian clock. *Science* **271**, 1736–1740 (1996).
19. Zeng, H. K., Qian, Z. W., Myers, M. P. & Rosbash, M. A light-entrainment mechanism for the *Drosophila* circadian clock. *Nature* **380**, 129–135 (1996).
20. Kaneko, M., Park, J. H., Cheng, Y., Hardin, P. E. & Hall, J. C. Disruption of synaptic transmission or clock-gene-product oscillations in circadian pacemaker cells of *Drosophila* cause abnormal behavioral rhythms. *J. Neurobiol.* **43**, 207–233 (2000).
21. Yang, Z. & Sehgal, A. Role of molecular oscillations in generating behavioral rhythms in *Drosophila*. *Neuron* **29**, 453–467 (2001).
22. Hunter-Ensor, M., Ousley, A. & Sehgal, A. Regulation of the *Drosophila* protein timeless suggests a mechanism for resetting the circadian clock by light. *Cell* **84**, 677–685 (1996).
23. Lee, C. G., Parikh, V., Itskovich, T., Bae, K. & Edery, I. Resetting the *Drosophila* clock by photic regulation of PER and a PER-TIM complex. *Science* **271**, 1740–1744 (1996).
24. Naidoo, N., Song, W., Hunter-Ensor, M. & Sehgal, A. A role for the proteasome in the light response of the Timeless clock protein. *Science* **285**, 1737–1741 (1999).
25. Price, M. A. & Kalderon, D. Proteolysis of the Hedgehog signaling effector Cubitus interruptus requires phosphorylation by glycogen synthase kinase 3 and casein kinase 1. *Cell* **108**, 823–835 (2002).
26. Jia, J. *et al.* Shaggy/GSK3 antagonizes Hedgehog signalling by regulating Cubitus interruptus. *Nature* **416**, 548–552 (2002).
27. Pai, L. M., Orsulic, S., Bejsovic, A. & Peifer, M. Negative regulation of Armadillo, a Wingless effector in *Drosophila*. *Development* **124**, 2255–2266 (1997).
28. Stanewsky, R. *et al.* Temporal and spatial expression patterns of transgenes containing increasing amounts of the *Drosophila* clock gene period and a lacZ reporter: Mapping elements of the PER protein involved in circadian cycling. *J. Neurosci.* **17**, 676–696 (1997).
29. Ruel, L., Pantescio, V., Lutz, Y., Simpson, P. & Bourouis, M. Functional significance of a family of protein kinases encoded at the shaggy locus in *Drosophila*. *EMBO J.* **12**, 1657–1669 (1993).
30. Cegielska, A., Gietzen, K. F., Rivers, A. & Virshup, D. M. Autoinhibition of casein kinase 1 ϵ (CK1 ϵ) is relieved by protein phosphatases and limited proteolysis. *J. Biol. Chem.* **273**, 1357–1364 (1998).

Supplementary Information accompanies the paper on Nature's website
 (http://www.nature.com/nature).

Acknowledgements We thank M. Boudinot for the Faas software, M. Serrier and L. Collet for help with the figures, M. Rosbash, P. Emery, A. Klarsfeld, J.-F. Julien and E. Petrochilo for their comments and suggestions on the manuscript, as well as J. Champagnat and J.-D. Vincent for their continuous support. We thank I. Miletich for the unpublished UAS-*slimb* line, and R. Myers, L. Saez, R. Stanewsky and D. Virshup for providing antibodies or constructs. This work was supported by CNRS (ATIP "Développement" and appel d'offres "Biologie cellulaire") and Fondation pour la Recherche Médicale. F.R. is supported by INSERM.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to F.R.
 (e-mail: rouyer@iaf.cnrs-gif.fr).

Non-redundant role of the long pentraxin PTX3 in anti-fungal innate immune response

Cecilia Garlanda^{*†}, Emilio Hirsch^{†‡}, Silvia Bozza[§], Antonietta Salustri^{||}, Marika De Acetis[‡], Rachele Nota^{*}, Alessia Maccagno^{||}, Federica Riva^{*}, Barbara Bottazzi^{*}, Giuseppe Peri^{*}, Andrea Doni^{*}, Luca Vago[¶], Marina Botto[#], Rita De Santis[☆], Paolo Carminati[☆], Gregorio Siracusa^{||}, Fiorella Altruda[‡], Annunziata Vecchi^{*}, Luigina Romani[§] & Alberto Mantovani^{*††}

^{*} Department of Immunology and Cell Biology, Mario Negri Institute for Pharmacological Research, 20157 Milan, Italy

[‡] Department of Genetics, Biology and Biochemistry, University of Turin, Turin, Italy

[§] Microbiology Section, Department of Experimental Medicine and Biochemical Sciences, University of Perugia, Perugia, Italy

^{||} Department of Public Health and Cell Biology, Roma Tor Vergata University, Rome, Italy

[¶] Institute of Pathology, University of Milan, Ospedale Luigi Sacco, Milan, Italy

[#] Rheumatology Section, Division of Medicine, Imperial College School of Medicine, Hammersmith Hospital, London, UK

[☆] SigmaTau SpA, Pomezia, Rome, Italy

^{††} Centro IDET, Institute of General Pathology, University of Milan, Milan, Italy

[†] These authors contributed equally to this work

Pentraxins are a superfamily of conserved proteins that are characterized by a cyclic multimeric structure¹. The classical short pentraxins, C-reactive protein (CRP) and serum amyloid P component (SAP), are acute-phase proteins produced in the liver in response to inflammatory mediators^{2–4}. Short pentraxins regulate innate resistance to microbes and the scavenging of cellular debris and extracellular matrix components^{2–5}. In contrast, long pentraxins have an unrelated, long amino-terminal domain coupled to the carboxy-terminal pentraxin domain, and differ, with respect to short pentraxins, in their gene organization, chromosomal localization, cellular source, and in their stimuli-inducing and ligand-recognition ability⁶. To investigate the *in vivo* function of the long pentraxin PTX3, we generated mice deficient in Ptx3 by homologous recombination. Ptx3-null mice were susceptible to invasive pulmonary aspergillosis. Ptx3 binds selected microbial agents, including conidia of *Aspergillus fumigatus*, and we found that susceptibility of Ptx3-null mice was associated with defective recognition of conidia by alveolar macrophages and dendritic cells, as well as inappropriate induction of an adaptive type 2 response. Thus, the long pentraxin Ptx3 is a secreted pattern-recognition receptor that has a non-redundant role in resistance to selected microbial agents, in particular to the opportunistic fungal pathogen *Aspergillus fumigatus*.

PTX3 is produced as a 10–20-subunit multimer protein by macrophages and other cell types or tissues on stimulation with primary inflammatory mediators, such as lipopolysaccharide, interleukin-1 and tumour-necrosis factor- α (LPS, IL-1 and TNF- α , respectively)^{7–13}. PTX3 does not recognize classical pentraxin ligands, but it binds to the complement component C1q (ref. 11) and selected microorganisms (see below). To assess the *in vivo* role of Ptx3, we generated *Ptx3*^{−/−} mice by homologous recombination (see Supplementary Information). As discussed elsewhere (our own unpublished data, see also ref. 14) *Ptx3*^{−/−} mice show a severe defect in female fertility. PTX3 binds selected pathogens, including *Pseudomonas aeruginosa*, *Salmonella typhimurium* and *Aspergillus fumigatus* (Fig. 1). Among the microbes recognized by PTX3, we focused our attention on *A. fumigatus* because this fungus is a major pathogen in immunodeficient individuals, and pattern-recognition

receptors that have a non-redundant role *in vivo* against fungal infections of mammals have not yet been described¹⁵. As shown in Fig. 1a–d, Ptx3 bound to viable or heat-inactivated *A. fumigatus* conidia, as assessed using biotin-labelled Ptx3 or anti-Ptx3 monoclonal antibody, with an apparent dissociation constant (k_d) of 7.8×10^{-7} M. Ptx3 also bound *Aspergillus flavus* and *Aspergillus niger*. The binding of Ptx3 to conidia was abolished in the presence of galactomannan, a major constituent of the conidium wall, but not in the presence of dextran, galactose, fucose or mannose. In contrast, Ptx3 did not bind to the hyphae of *A. fumigatus*. Similarly, Ptx3 did not bind to the unrelated fungus *Candida albicans*.

We then investigated whether Ptx3 affects the interaction of conidia with macrophages, which have a principal role in resistance to *A. fumigatus*¹⁶. Ptx3 facilitates the ingestion of conidia by macrophages, an effect more marked in *Ptx3*^{-/-} mice (Fig. 2a). Moreover, an increase in monocyte chemotactic protein-1 (MCP-1, also known as CCL-2) was observed in mononuclear phagocytes exposed to conidia and Ptx3, compared with conidia alone (Fig. 2b). Under the same conditions, Ptx3 did not significantly affect activation by LPS (Fig. 2b).

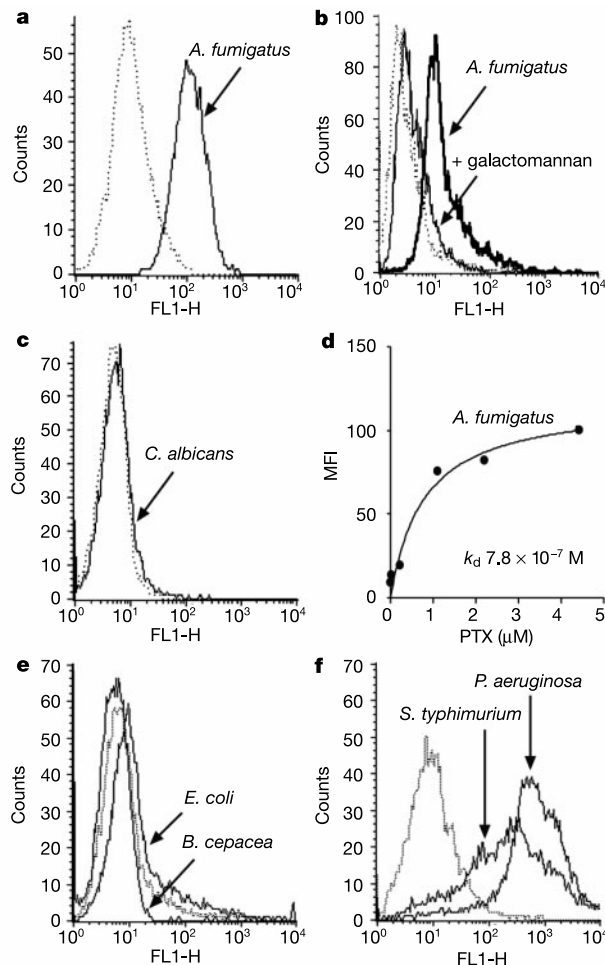


Figure 1 Interaction of Ptx3 with selected pathogens. **a–c**, FACS analysis of Ptx3 binding to conidia of *A. fumigatus* and *C. albicans*. Binding was revealed by biotinylated Ptx3 followed by fluorescein isothiocyanate (FITC)-labelled streptavidin. Heat-inactivated (**a**) or viable (**b**) conidia were used with similar results. Galactomannan ($20 \mu\text{g ml}^{-1}$) abolished binding (**b**). **d**, Specific binding of Ptx3 to *A. fumigatus* conidia. Binding is saturable with a k_d of 7.8×10^{-7} M (calculated on the basis of the Ptx3 protomer mass of 45 kDa). MFI, mean fluorescence intensity. **e, f**, Binding to bacteria. Dotted curves show control fluorescence in the absence of Ptx3.

Having shown that Ptx3 can bind conidia and facilitate their interaction with mononuclear phagocytes, we next assessed whether conidia were able to induce production of Ptx3, as also suggested by the defective recognition of conidia by *Ptx3*^{-/-} macrophages. Conidia induced PTX3 production in human and murine mononuclear phagocytes and dendritic cells, with modest or no induction in endothelial cells, epithelial cells, fibroblasts and neutrophils (Fig. 2c, and data not shown). Induction was rapid, with immunoreactive Ptx3 being detectable in cells after 1 h (Fig. 2d). As expected on the basis of these *in vitro* results, infection of mice and humans with *A. fumigatus* results in induction of PTX3 in bronchoalveolar lavage fluids and plasma. For instance, 24 h after intratracheal (i.t.) challenge with conidia, Ptx3 plasma levels were 128.2 ng ml^{-1} (range 103.3–153.2) compared with 9.92 ng ml^{-1} (range 1.85–37.74) for control mice. Depletion of neutrophils by treatment with cyclophosphamide or antibody (Table 1) did not affect Ptx3 induction. The bronchoalveolar lavage fluid (1 ml of saline) contained $5.01 \pm 4 \text{ ng ml}^{-1}$ Ptx3, whereas Ptx3 was undetectable in normal animals. Finally, ten neutropenic patients with haematological malignancies and *A. fumigatus* systemic infection had PTX3 plasma levels significantly increased compared with control subjects (9.56 ± 3.15 and $1.08 \pm 0.13 \text{ ng ml}^{-1}$ for infected and control subjects, respectively; $P < 0.0001$). Moreover, immunohistochemistry of lungs from a patient with invasive pulmonary aspergillosis demonstrated a diffuse cytoplasmic positivity in alveolar macrophages and in a few circulating mononuclear cells, whereas epithelial and stromal cells were negative (Fig. 2e).

The finding that Ptx3 facilitated recognition of conidia by macrophages implied the existence of a receptor for this molecule. As shown in Fig. 2f, Ptx3 bound to mononuclear phagocytes and dendritic cells, but not to T lymphocytes.

As Ptx3 bound *A. fumigatus* conidia *in vitro* and facilitated their interaction with macrophages, we assessed a potential role for Ptx3 in resistance in a murine model of invasive pulmonary aspergillosis (IPA)¹⁷. As shown in Table 1 and Fig. 3a, all wild-type mice survived infection. In contrast, all *Ptx3*^{-/-} mice died with a median survival time (MST) of 3 days (two experiments performed). The defective resistance of *Ptx3*^{-/-} mice was equally evident in the 129/Sv and in the 129/Sv $\times$ C57BL/6 background (one experiment not shown). The increased susceptibility of *Ptx3*^{-/-} mice correlated with a marked increase in lung and brain colonization (Table 1). Lungs

Table 1 Susceptibility of *Ptx3*^{-/-} mice to invasive pulmonary aspergillosis

Mice	Treatment	MST (days)	Dead/total	Brain CFU†	Lung CFU‡
Experiment 1					
<i>Ptx3</i> ^{+/+}	None	>60	0/3	0	8,100
<i>Ptx3</i> ^{+/+}	RB6-8C5	4	4/4	34,800	170,100
<i>Ptx3</i> ^{-/-}	None	3	3/3	142,200	706,500
<i>Ptx3</i> ^{-/-}	RB6-8C5	3	3/3	187,200	603,750
Experiment 2					
<i>Ptx3</i> ^{+/+}	None	>60	0/6	ND	12,900
<i>Ptx3</i> ^{-/-}	None	3	7/7	ND	233,250
<i>Ptx3</i> ^{-/-}	PTX3*	>60	0/4	ND	60,900
<i>Ptx3</i> ^{-/-}	PTX3†	>60	1/4	ND	101,700
Experiment 3					
BMT§	None	3	8/8	ND	814,310
BMT§	PTX3*	8	6/8	ND	187,300
Experiment 4					
<i>C1q</i> ^{+/+}	None	>60	0/10	0	1,800
<i>C1q</i> ^{-/-}	None	2	10/10	450	400,000
<i>C1q</i> ^{-/-}	PTX3*	>60	0/6	ND	78,430

Mice were infected i.t. with *A. fumigatus* conidia (2×10^6 per mouse) on day 0. Neutrophil depletion was obtained by treatment with RB6-8C5 monoclonal antibody ($100 \mu\text{g}$ per mouse) i.p. 2 h before fungal challenge. MST, median survival time; ND, not determined.

*PTX3 was administered on day 0 i.t. and on days 1 and 2 i.v. at $20 \mu\text{g}$ per mouse.

†PTX3 administered on day 0 i.t. at a dose of $20 \mu\text{g}$ per mouse and on days 1 and 2 i.p. at $50 \mu\text{g}$ per mouse.

‡CFU were determined on day 3 after infection.

§Mice underwent allogeneic T-cell-depleted BMT as described¹⁸ and were infected i.t. with *A. fumigatus* conidia (2×10^6 per mouse) 7 days later.

|| $P < 0.05$ compared with control mice (Mann–Whitney U-test).

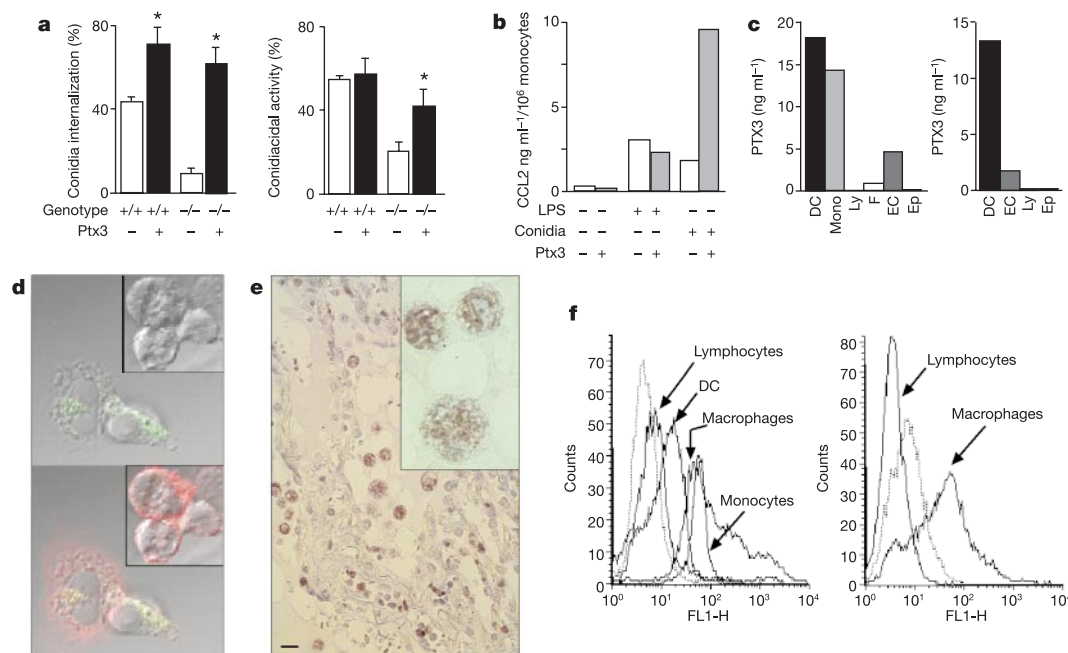


Figure 2 Induction of Ptx3 by *A. fumigatus* *in vitro* and *in vivo* and its role in promoting recognition of conidia. **a**, Phagocytosis and killing of *A. fumigatus* conidia by alveolar macrophages in the absence (open columns) or presence (closed columns) of Ptx3. Asterisks indicate $P < 0.05$ (Student's *t*-test). **b**, Amplification by Ptx3 ($2 \mu\text{g ml}^{-1}$) of conidia-induced CCL2 production. **c**, Induction of Ptx3 by *A. fumigatus* conidia in human (left) and murine (right) mononuclear phagocytes and dendritic cells. DC, dendritic cells; EC, endothelial cells; Ep, epithelial cells; F, fibroblasts; Ly, T lymphocytes; Mono, monocyte/macrophages. **d**, Rapid production of Ptx3 in response

to conidia. Confocal microscopy of immunoreactive Ptx3 (green fluorescence) in murine macrophages (CD11b⁺, red fluorescence) exposed to conidia for 1 h is shown. Unstimulated cells show no staining (insets). **e**, Immunohistochemical demonstration of Ptx3 in a clinical case of IPA. Several alveolar macrophages (inset) and a few circulating cells show cytoplasmic positivity for Ptx3 (brown). Other cell types are negative. Scale bar, 20 μm . **f**, Binding of biotin-labelled Ptx3 ($2.2 \times 10^{-6} \text{ M}$) to murine (right) and human (left) mononuclear phagocytes and dendritic cells. Dotted curves show control fluorescence with irrelevant antibodies.

from infected *Ptx3*^{-/-} mice showed a massive inflammatory response, the presence of hyphae and numerous, mainly extracellular, conidia (Fig. 3b, top row), as opposed to the few intracellular conidia and signs of a modest inflammatory reaction observed in controls (Fig. 3b, bottom row). In two other experiments *Ptx3*^{-/-} mice were treated with purified Ptx3 i.t. at the time of challenge (day 0) and intravenously or intraperitoneally on day 1 and 2 with protection against IPA (Table 1). Mock control preparations were inactive.

Aspergillus fumigatus is a major opportunistic pathogen in immunodeficient patients and poses a formidable therapeutic challenge¹⁶. We therefore investigated whether administration of PTX3 was active in an IPA model of allogeneic, T-cell-depleted, bone marrow transplantation (BMT)¹⁸. As shown in Table 1, combined systemic and local PTX3 administration caused a significant twofold increase in survival time with two out of eight mice being cured. Moreover, the lung colony-forming unit (CFU) counts were markedly reduced (over fourfold) in PTX3-treated mice.

It was important to assess whether the susceptibility of *Ptx3*^{-/-} mice to *A. fumigatus* reflected a generalized impairment of host resistance to microbial pathogens. *Ptx3*^{-/-} mice showed accelerated death when challenged with *P. aeruginosa*, a pathogen recognized by Ptx3 (Fig. 1), with a tenfold increase in lung CFU at 24 h (Fig. 3a). In contrast, susceptibility to *Listeria monocytogenes* and to intra-abdominal sepsis caused by caecal ligation and puncture were not affected by Ptx3 deficiency (Fig. 3a; see also Supplementary Information).

Resistance to *A. fumigatus* requires phagocytes and is associated with the activation of a polarized type 1 T-cell response^{19–21}. *In vitro*, the ability of alveolar macrophages to ingest and kill resting conidia was impaired in *Ptx3*^{-/-} mice, as compared with *Ptx3*^{+/+} mice

(Fig. 2a). Addition of Ptx3 restored both the phagocytic and conidiocidal activities of cells from *Ptx3*^{-/-} mice and, to a lesser extent, potentiated those of *Ptx3*^{+/+} mice, as discussed above (Fig. 2a). Interferon- γ (IFN- γ) and IL-12 levels were reduced and IL-4 levels increased in lungs of *Ptx3*^{-/-} mice (Fig. 3c). This defect was reverted by the administration of PTX3. These results suggest that in *Ptx3*^{-/-} mice, the defective recognition of *A. fumigatus* conidia by the host is associated with the lack of development of appropriate T-helper-cell type 1 (T_H1)-type responses and to an unbalanced cytokine profile skewed towards a T_H2 response.

Lung dendritic cells are important in driving T-cell polarization in response to *A. fumigatus in vivo*²². Activation by interaction with conidia (assessed as production of IL-12, and upregulation of major histocompatibility complex class II (MHC II) and CD86 antigens) was defective in dendritic cells from *Ptx3*^{-/-} mice (Fig. 3). Addition of Ptx3 fully restored the ability of *Ptx3*^{-/-} dendritic cells to respond to conidia. *Ptx3*^{-/-} dendritic cells showed normal responsiveness to LPS (data not shown). These results suggest that by favouring recognition of conidia, Ptx3 facilitates macrophage-mediated resistance against *A. fumigatus* as well as activation by dendritic cells of a protective type 1 anti-fungal response.

Ptx3 binds to C1q and activates the complement cascade¹¹. *C1q*^{-/-} mice²³ showed increased susceptibility to IPA (Table 1). Treatment with PTX3 reverted this susceptibility (Table 1). Therefore Ptx3 can mediate resistance against *A. fumigatus* independently of C1q.

Our results demonstrate that Ptx3 acts as a soluble pattern-recognition receptor, which has a non-redundant role in resistance against the fungal pathogen *A. fumigatus*. We are not aware of previous unequivocal evidence for an essential role of soluble or membrane pattern-recognition receptors in resistance to fungi in

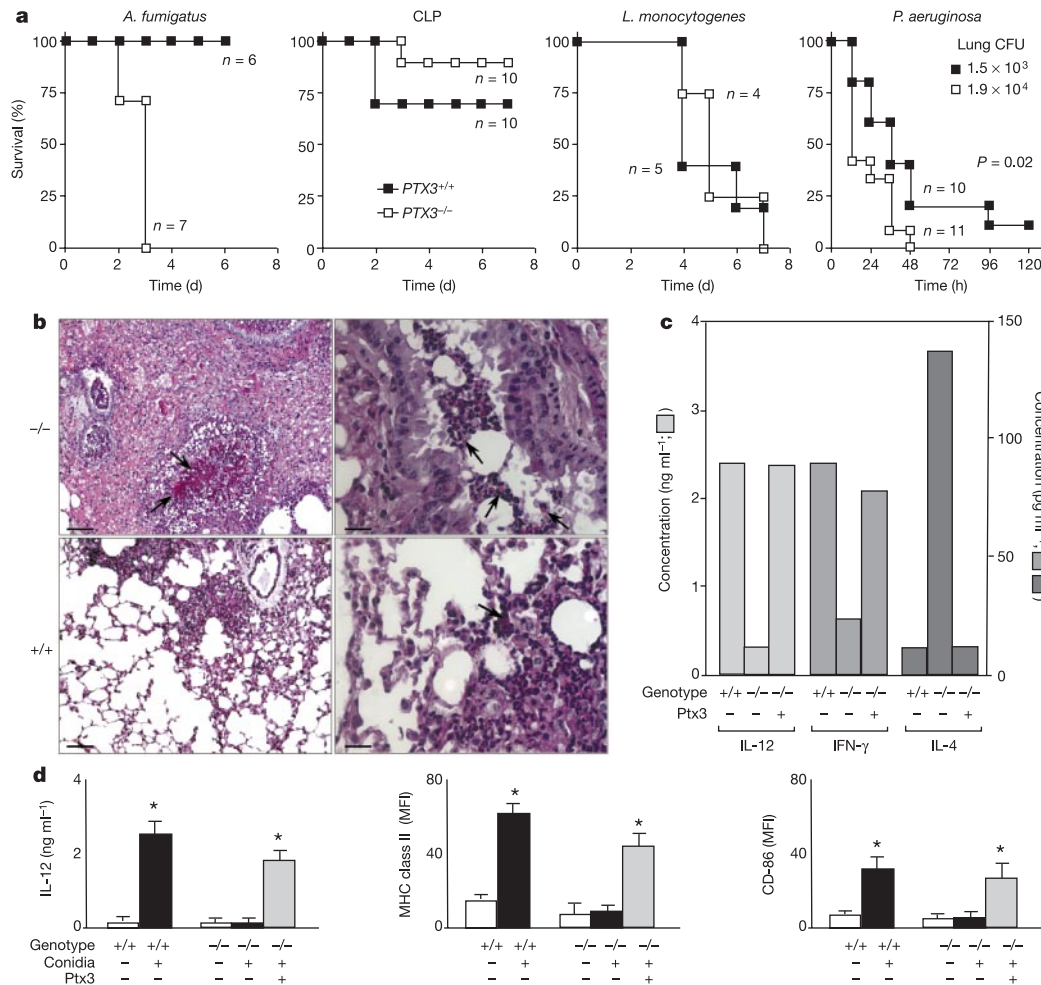


Figure 3 Selective susceptibility of *Ptx3*^{-/-} mice to *A. fumigatus*. **a**, Survival of *Ptx3*^{-/-} (open squares) and *Ptx3*^{+/+} (closed squares) mice. The inset in the far right panel shows the number of CFU per lung at 24 h. Results are representative of two to four experiments performed with three to ten mice per group (see also Supplementary Information) except for *P. aeruginosa* (one experiment). CLP, caecal ligation and puncture. **b**, Histopathology of lungs with IPA (day 2). Scale bar: 200 μm , first column; 50 μm ,

second column. Arrows point to conidia. **c**, Polarization of cytokine production in the lungs of *Ptx3*^{-/-} mice with IPA (day 3). PTX3 (20 μg per mouse i.t. and i.p.) was given on days 0, 1 and 2. One experiment is shown, representative of three performed. **d**, Role of Ptx3 in the activation of lung dendritic cells by conidia. MFI, mean fluorescence intensity. Asterisks indicate $P < 0.05$ (Student's *t*-test).

mammals. The regulated expression of this molecule in macrophages and dendritic cells suggests that Ptx3, unlike the short pentraxins made in the liver, represents a mechanism of amplification of innate resistance against pathogens mainly acting locally at sites of infection and inflammation. The therapeutic potential of Ptx3 in BMT and similar contexts requires further analysis. Ptx3 is unique in having a non-redundant, dual function *in vivo* in the control of female fertility (our own unpublished data, and ref. 14) and pathogen resistance.

Methods

Generation of *Ptx3*^{-/-} mice

For a detailed description, see Supplementary Information. Phenotypic analysis was carried out on the two lines derived from independent clones, and results were confirmed in a 129/Sv-C57Bl/6 mixed and 129/Sv inbred genetic background.

Ptx3 messenger RNA and protein

Northern blot analysis, expression and purification of Ptx3, and antibody assays were performed as described^{10,11,24}. Ptx3 was expressed in Chinese hamster ovary (CHO) cells and purified under endotoxin-free conditions by immunoaffinity²⁵ (Supplementary Information).

Cells

Alveolar and peritoneal macrophages, lung dendritic cells, epithelial cells and endothelial cells were isolated as described^{7,10,22}. Human monocytes were isolated from peripheral blood. Human macrophages and dendritic cells were generated from monocytes (Supplementary Information).

Ptx3 binding

Binding of Ptx3 to fungi, bacteria or eukaryotic cells was characterized using biotin-labelled Ptx3 at concentrations ranging from 1 to 200 $\mu\text{g ml}^{-1}$ ($(0.02\text{--}4.4) \times 10^{-6}\text{M}$ assuming a molecular mass of 45 kDa for the Ptx3 protomer) or anti-Ptx3 monoclonal antibody. For binding inhibition, 20 $\mu\text{g ml}^{-1}$ galactomannan, mannose, fucose, dextran or galactose (Sigma) were added to a mixture of conidia and biotin-labelled Ptx3. Binding was evaluated by fluorescence-activated cell sorting (FACS). The dissociation constant was calculated by nonlinear fitting of the saturation curves using the mean channel of fluorescence. Data are the mean of three experiments.

Infections

Invasive pulmonary aspergillosis (IPA) was studied by i.t. injection of 2×10^8 conidia as described¹⁷. Caecal ligation and puncture, and *L. monocytogenes* infection, were performed as described^{26,27}. *Pseudomonas aeruginosa* pulmonary infection was performed by i.t. injection of 10^7 bacteria (Supplementary Information).

Cytokine assays

The levels of IFN- γ , IL-4, CCL2 and IL-12 in supernatants or lung homogenate were determined by cytokine-specific enzyme-linked immunosorbent assays, as described¹⁷.

Phagocytosis and conidiocidal assays

Alveolar macrophages were exposed to conidia at a macrophage:conidia ratio of 1:5 for 2 h, and 10:1 for 4 h, in the absence or presence of $20 \mu\text{g ml}^{-1}$ Ptx3 (0.44×10^{-6} M Ptx3 protomer) before being evaluated for internalization or conidiocidal activity, respectively (Supplementary Information).

Received 26 July; accepted 8 October 2002; doi:10.1038/nature01195.

1. Emsley, J. *et al.* Structure of pentameric human serum amyloid P component. *Nature* **367**, 338–345 (1994).
2. Szalai, A. J., Agrawal, A., Greenhough, T. J. & Volanakis, J. E. C-reactive protein structural biology, gene expression, and host defense function. *Immunol. Res.* **16**, 127–136 (1997).
3. Steel, D. M. & Whitehead, A. S. The major acute phase reactants: C-reactive protein: serum amyloid P component and serum amyloid A protein. *Immunol. Today* **15**, 81–88 (1994).
4. Pepys, M. B. & Baltz, M. L. Acute phase proteins with special reference to C-reactive protein and related proteins (pentaxins) and serum amyloid A protein. *Adv. Immunol.* **34**, 141–212 (1983).
5. Noursadeghi, M. *et al.* Role of serum amyloid P component in bacterial infection: protection of the host or protection of the pathogen. *Proc. Natl Acad. Sci. USA* **97**, 14584–14589 (2000).
6. Mantovani, A., Garlanda, C. & Bottazzi, B. *Cytokine Reference* (eds Oppenheim, J. & Feldmann, M.) (Academic, New York, 2002).
7. Breviaro, F. *et al.* Interleukin-1-inducible genes in endothelial cells. Cloning of a new gene related to C-reactive protein and serum amyloid P component. *J. Biol. Chem.* **267**, 22190–22197 (1992).
8. Lee, G. W., Lee, T. H. & Vilcek, J. TSG-14, a tumour necrosis factor- and IL-1-inducible protein, is a novel member of the pentaxin family of acute phase proteins. *J. Immunol.* **150**, 1804–1812 (1993).
9. Vidal-Alles, V. *et al.* Inducible expression of PTX3, a new member of the pentraxin family, in human mononuclear phagocytes. *Blood* **84**, 3483–3493 (1994).
10. Introna, M. *et al.* Cloning of mouse PTX3, a new member of the pentraxin gene family expressed at extrahepatic sites. *Blood* **87**, 1862–1872 (1996).
11. Bottazzi, B. *et al.* Multimer formation and ligand recognition by the long pentraxin PTX3 – similarities and differences with the short pentraxins C-reactive protein and serum amyloid P component. *J. Biol. Chem.* **272**, 32817–32823 (1997).
12. Rovere, P. *et al.* The long pentraxin PTX3 binds to apoptotic cells and regulates their clearance by antigen-presenting dendritic cells. *Blood* **96**, 4300–4306 (2000).
13. Lee, G. W., Goodman, A. R., Lee, T. H. & Vilcek, J. Relationship of TSG-14 protein to the pentraxin family of major acute phase proteins. *J. Immunol.* **153**, 3700–3707 (1994).
14. Varani, S. *et al.* Knockout of pentraxin 3, a downstream target of growth differentiation factor-9, causes female subfertility. *Mol. Endocrinol.* **16**, 1154–1167 (2002).
15. Medzhitov, R. Toll-like receptors and innate immunity. *Nature Rev. Immunol.* **1**, 135–145 (2001).
16. Denning, D. W. Invasive aspergillosis. *Clin. Infect. Dis.* **26**, 781–803 (1998).
17. Cenci, E. *et al.* Th1 and Th2 cytokines in mice with invasive aspergillosis. *Infect. Immun.* **65**, 564–570 (1997).
18. Menacacci, A. *et al.* Defective antifungal T-helper 1 (TH1) immunity in a murine model of allogeneic T-cell-depleted bone marrow transplantation and its restoration by treatment with TH2 cytokine antagonists. *Blood* **97**, 1483–1490 (2001).
19. Schaffner, A., Douglas, H. & Braude, A. Selective protection against conidia by mononuclear and against mycelia by polymorphonuclear phagocytes in resistance to *Aspergillus*. Observations on these two lines of defense *in vivo* and *in vitro* with human and mouse phagocytes. *J. Clin. Invest.* **69**, 617–631 (1982).
20. Romani, L. The T cell response against fungal infections. *Curr. Opin. Immunol.* **9**, 484–490 (1997).
21. Menacacci, A. *et al.* Cytokines in candidiasis and aspergillosis. *Curr. Pharm. Biotechnol.* **1**, 235–251 (2000).
22. Bozza, S. *et al.* Dendritic cells transport conidia and hyphae of *Aspergillus fumigatus* from the airways to the draining lymph nodes and initiate disparate Th responses to the fungus. *J. Immunol.* **168**, 1362–1371 (2002).
23. Botto, M. *et al.* Homozygous C1q deficiency causes glomerulonephritis associated with multiple apoptotic bodies. *Nature Genet.* **19**, 56–59 (1998).
24. Peri, G. *et al.* PTX3, a prototypic long pentraxin, is an early indicator of acute myocardial infarction in man. *Circulation* **102**, 636–641 (2000).
25. Muller, B. *et al.* Circulating levels of the long pentraxin PTX3 correlate with severity of infection in critically ill patients. *Crit. Care Med.* **29**, 1404–1407 (2001).
26. Villa, P. *et al.* Pattern of cytokines and pharmacomodulation in sepsis induced by cecal ligation and puncture compared with that induced by endotoxin. *Clin. Diagn. Lab. Immunol.* **2**, 549–553 (1995).
27. Watanabe, Y., Mitsuyama, M., Sano, M., Nakano, H. & Nomoto, K. Enhanced resistance against *Listeria monocytogenes* at an early phase of primary infection in pregnant mice: activation of macrophages during pregnancy. *Infect. Immun.* **52**, 730–735 (1986).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements This work was supported by Istituto Superiore di Sanità, Ministero Istruzione, Università e Ricerca (MIUR), Consiglio Nazionale della Ricerche (CNR), and by the European Commission. We acknowledge the contribution of the Italian Association for Cancer Research. We thank C. Scotton for critical reading of the manuscript.

Competing interests statement The authors declare competing financial interests: details accompany the paper on Nature's website (<http://www.nature.com/nature>).

Correspondence and requests for materials should be addressed to A.M. (e-mail: mantovani@marionegri.it).

Escherichia coli K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth

Rafael U. Ibarra^{*†}, Jeremy S. Edwards^{†‡} & Bernhard O. Palsson^{*}

^{*} Department of Bioengineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093-0412, USA

[‡] Department of Chemical Engineering, University of Delaware, Newark, Delaware 19716, USA

[†] These authors contributed equally to this work

Annotated genome sequences^{1,2} can be used to reconstruct whole-cell metabolic networks^{3–6}. These metabolic networks can be modelled and analysed (computed) to study complex biological functions^{7–11}. In particular, constraints-based *in silico* models¹² have been used to calculate optimal growth rates on common carbon substrates, and the results were found to be consistent with experimental data under many but not all conditions^{13,14}. Optimal biological functions are acquired through an evolutionary process. Thus, incorrect predictions of *in silico* models based on optimal performance criteria may be due to incomplete adaptive evolution under the conditions examined. *Escherichia coli* K-12 MG1655 grows sub-optimally on glycerol as the sole carbon source. Here we show that when placed under growth selection pressure, the growth rate of *E. coli* on glycerol reproducibly evolved over 40 days, or about 700 generations, from a sub-optimal value to the optimal growth rate predicted from a whole-cell *in silico* model. These results open the possibility of using adaptive evolution of entire metabolic networks to realize metabolic states that have been determined a priori based on *in silico* analysis.

Predictive whole-cell metabolic models can be developed using a constraints-based modelling procedure^{15–18}. As an alternative to detailed theory-based models, constraints-based models use the successive imposition of governing constraints (such as mass conservation, thermodynamics, capacity and nutritional environment) to eliminate network functions that exceed the governing constraints. Mathematically this procedure defines a solution space containing all possible metabolic network functions that satisfy the governing constraints. Each particular solution in this space corresponds to a particular state of the metabolic network and therefore a potential behaviour of the cell. Within the solution space defined by the governing constraints, the optimal use of the metabolic network to support growth can be found among all possible solutions using linear optimization^{16–19}. However, a single optimal growth condition is of limited interest and a phenotype phase plane (PPP) analysis has been developed to obtain a broad understanding of a metabolic network's optimal properties^{20,21}. The PPP analysis evaluates the optimal properties of a metabolic network under a range of environmental conditions (see Methods) and has been used to show that the growth of *E. coli* is consistent with the optimal use of its metabolic network under several defined growth conditions^{12–14}.

It is not known whether optimal growth is observed on all substrates, and if not, whether adaptive evolution towards optimal growth can be achieved. Furthermore, if such adaptive evolution towards the optimal behaviour occurs, does the endpoint correspond with a priori calculations? To address these issues, we examined prolonged exponential growth of *E. coli* K-12 on several substrates (acetate, succinate, malate, glucose and glycerol). All calculations presented here were made with a previously formulated large-scale *E. coli* metabolic model^{12,14}, and the model was not adjusted or 'fitted' to the data described.

Batch growth experiments were done using malate as the sole

carbon source with a range of substrate concentrations ($0.25\text{--}3\text{ g l}^{-1}$) and temperatures ($29\text{--}37^\circ\text{C}$) to vary the malate uptake rate (MUR). The MUR, oxygen uptake rate (OUR) and growth rate were measured. The measured MUR and OUR data were optimal, as defined by the line of optimality (LO) in the PPP (Fig. 1a). The optimal growth rate of *E. coli* was calculated for all combinations of the MUR and OUR and displayed as a surface over the PPP (Fig. 1b). The experimentally determined growth rates were on the edge of the colour-coded solution space that corresponds to the LO (Fig. 1b). Hence the optimal growth performance of *E. coli* K-12 on malate was predicted *a priori* by using PPP. The results for growth with malate as the sole carbon source were in agreement with previous observations of *E. coli* metabolism for growth on succinate or acetate¹⁴.

A natural question arises: is the optimal performance on malate stable over prolonged periods of time? To address this question, adaptive evolution of *E. coli* on malate was studied for 500 generations. The adaptation resulted in a 19% increase in growth rate. However, the MUR and OUR also increased and maintained metabolic operation on the LO (Fig. 1). Similar adaptive evolution experiments on acetate and succinate resulted in an increased growth rate (20% and 17%, respectively) (Fig. 2). Both the oxygen and substrate uptake rates increased concomitantly to maintain optimal growth as defined and predicted by the PPP analysis.

The growth rate of *E. coli* using glucose as the carbon source was also increased by prolonged exponential growth (Figs 2 and 3). Before adaptive evolution on glucose the cellular growth rate, OUR, and glucose uptake rate (GUR) were experimentally determined over a range of glucose concentrations and temperatures. The experimentally determined values for the GUR and OUR corresponded to points on the LO or slightly in phase 2 (the acetate overflow region) of the PPP (ref. 21) (Fig. 3a). The predicted acetate secretion in phase 2 was experimentally observed and the measured growth rates were on the surface of the solution space near the edge corresponding to the LO (data not shown). *E. coli* was subsequently kept in sustained exponential growth over 500 generations (Fig. 3b, c). The growth rate increased by 17%, as shown by movement of the experimental data points within phase 2 on the surface towards higher growth rates. Thus, as with malate, succinate and acetate, the growth rate of *E. coli* with glucose as the carbon source could be slightly increased with the substrate and oxygen uptake rates moving in phase 2 with some acetate overflow. It was also noted that evolutionary adaptation maintained metabolic operation on the surface of the three-dimensional PPP, as predicted by the physicochemical constraints on the metabolic network. The metabolic operation in the phase 2 provided an increased growth rate

with a reduced yield (relative to the LO).

We determined the growth performance over a range of glycerol concentrations and temperature. Unlike growth on malate or glucose, the experimental data points were scattered throughout phase 1 far from the LO (Fig. 4b), indicating sub-optimal growth of

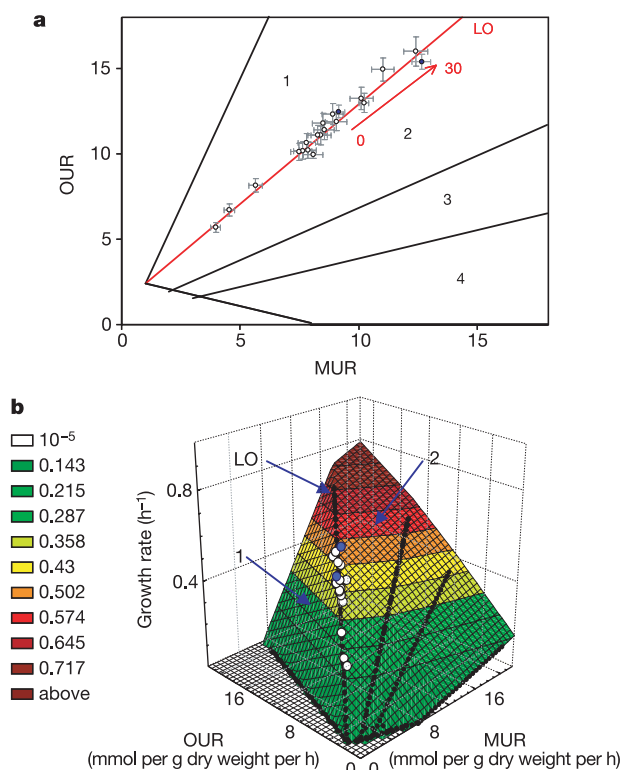


Figure 1 Growth of *E. coli* K-12 on malate. **a**, The malate–oxygen phenotype phase plane (PPP) Phase 1 is characterized by metabolic futile cycles, whereas phase 2 is characterized by acetate overflow metabolism. The line of optimality (LO, in red) separates phases 1 and 2 (ref. 21). Data points (open circles) represent malate concentrations ranging from $0.25\text{--}3\text{ g l}^{-1}$; and temperatures ranging from $29\text{--}37^\circ\text{C}$. The two data points in blue represent the starting point (day 0) and endpoint (day 30) of adaptive evolution respectively, at a malate concentration of 2 g l^{-1} and a temperature of 37°C . These data points represent a span of 500 generations. **b**, Three-dimensional representation of growth rates. The x and y axes represent the same variables as in **a**. The z axis represents the cellular growth rate (h^{-1}). OUR, oxygen uptake rate; MUR, malate uptake rate.

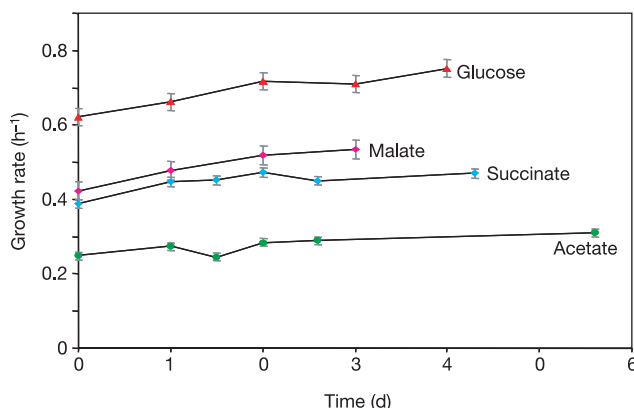


Figure 2 Growth rate during adaptive evolution on glucose, malate, succinate and acetate. Growth conditions were kept constant at a temperature of 37°C and a substrate concentration of 2 g l^{-1} . We measured growth rate in the exponential phase of growth.

The increases in growth rate over time were as follows: glucose (18%), malate (21%), succinate (17%) and acetate (20%). The number of generations for each adaptive evolution was: glucose (500), malate (500), succinate (1,000) and acetate (700).

wild-type *E. coli* K-12 on glycerol, consistent with previous observations²².

We studied *E. coli* adaptive growth using glycerol as the sole carbon source (2 g l^{-1}), by serial transfer, at a temperature of 30°C , and with sufficient oxygenation. Growth rate, glycerol uptake rate (GI-UR) and OUR were measured every ten days. Over a 40-day period an evolutionary path (E1) was observed (Fig. 4c). Phenotypic changes were traced in phase 1, eventually converging towards the LO. During this 40-day period, the growth rate more than doubled from 0.23 h^{-1} to 0.55 h^{-1} (Fig. 4a). The substrate uptake and growth rate data obtained under various growth conditions after adaptive evolution were near the LO (Fig. 4d). The evolved strain attained near-optimal growth on glycerol as defined by the *in silico* predictions. A second, independent adaptation experiment gave a similar but non-identical evolutionary trajectory (E2), converging near the same endpoint (Fig. 4c). Finally, a third independent adaptation experiment (E3) was done with a different initial starting point within phase 1. E3 was done at 37°C and a glycerol concentration of 2 g l^{-1} . The adaptation of *E. coli* to growth

on glycerol at 37°C resulted in motion towards the LO and the growth rate increased by about 30% (Fig. 4a). The final growth rate of the E3 strain was consistent with the *in silico* predictions with respect to the GI-UR, OUR and the growth rate.

To assess the stability of the endpoint of the adaptive evolution, we extended the cultivation on glycerol for an additional 300 generations, or 20 days for the E1 and E2 strains. The data indicated no further change in growth (Fig. 4a). On the sixtieth day of the experiment the E1 and E2 strains exhibited growth on the LO under various growth conditions, reaffirming optimal growth behaviour and the stability of the phenotype (Fig. 4e).

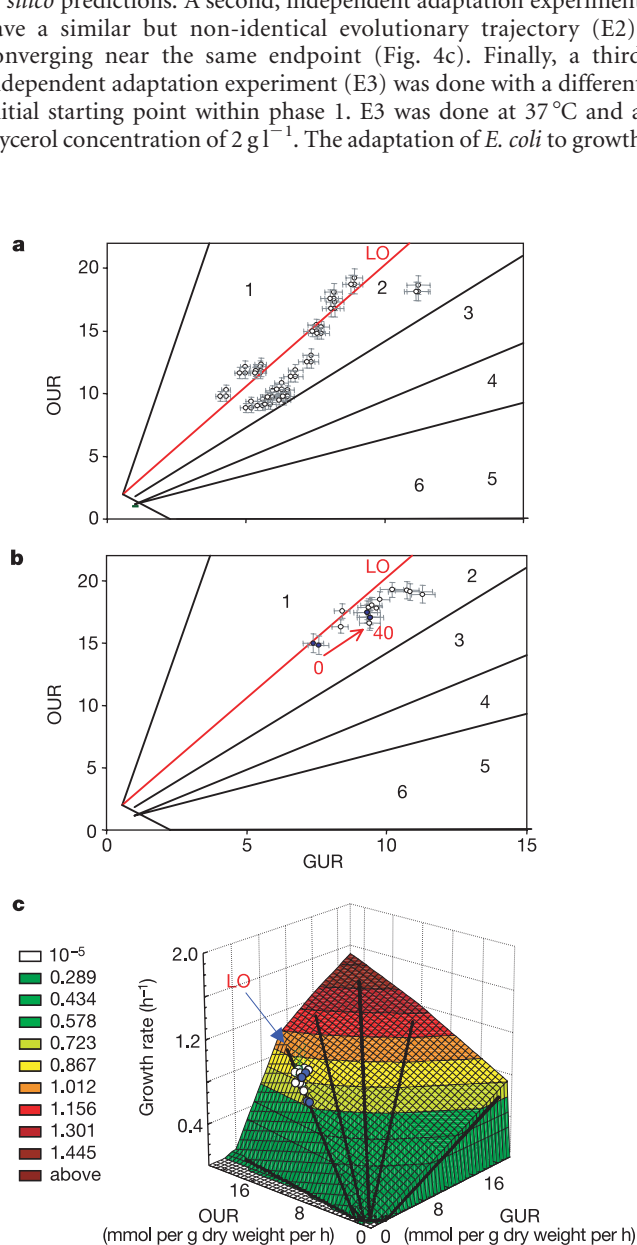


Figure 3 Growth of *E. coli* K-12 on glucose. **a**, The glucose–oxygen PPP. Like the malate–oxygen PPP, phase 1 represents sub-optimal growth and phase 2 is characterized by acetate overflow metabolism. The LO is shown in red. **b**, GUR plotted against OUR along with experimental values for adaptive evolution experiments. Open circles represent measurements during the adaptive evolutionary process, whereas blue circles indicate the beginning and end of evolution. **c**, Three-dimensional rendering of computed growth rate and the experimental data (from **a** and **b**). The *x* and *y* axes represent the GUR and OUR. The *z* axis represents the cellular growth rate (h^{-1}).

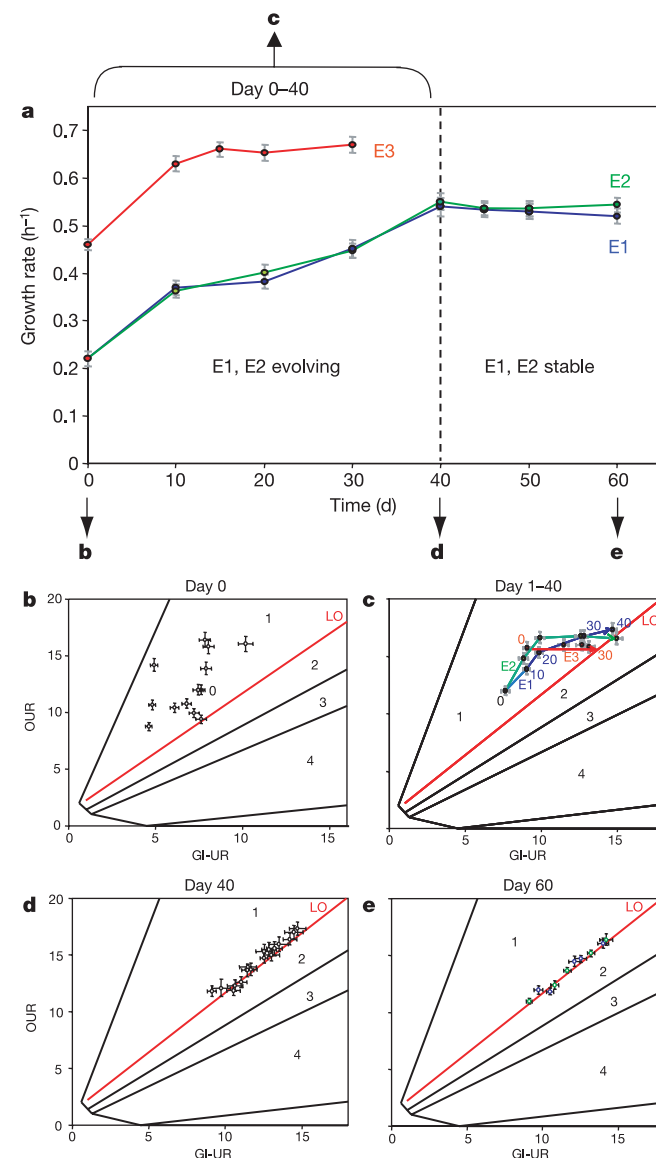


Figure 4 Growth of *E. coli* K-12 on glycerol. **a**, Change in growth rate with time for three adaptive evolution experiments: trajectories E1, E2, and E3. E1 and E2 were performed at 30°C and E3 at 37°C . The glycerol concentration was kept constant at 2 g l^{-1} during E1, E2 and E3. **b**, The PPP pre-evolution. The LO is shown in red. The range of glycerol concentrations used was $0.25\text{--}2 \text{ g l}^{-1}$. **c**, The PPP during adaptive evolution. Experimental values for E1 are indicated in blue, and for E2 they are indicated in green, and for E3 in red. The starting point of evolution for E1 and E2 is indicated in black (day 0). **d**, The PPP after 40 days (about 700 generations) of evolution. The range of glycerol concentrations used was $0.25\text{--}2 \text{ g l}^{-1}$. **e**, The PPP after 60 days (1,000 generations) of evolution. The range of glycerol concentrations used was $0.25\text{--}2 \text{ g l}^{-1}$. Data points were obtained using the E1 (blue) and E2 (green) strains. GI-UR, glycerol uptake rate.

Selection pressure is expected to result in optimal performance through an evolutionary process. Optimal growth of *E. coli* on acetate, succinate, malate and glucose is consistent with the predictions of whole-cell *in silico* models. The strain used here has presumably never had to compete for survival using glycerol as the sole carbon source and thus initially utilized this carbon source non-optimally. However, adaptive evolution on glycerol resulted in the *a priori* calculated optimal growth that was based on the constraints placed on the *E. coli* metabolic network. The adaptive evolutionary process had a reproducible and predictable endpoint.

This study opens up several possibilities. First, it may now be possible to specify optimal network properties *in silico* and achieve them through an adaptive evolutionary process or in combination with a series of other methodologies. *In silico* design of micro-organisms could be used to improve their metabolic abilities, production efficiency and/or operational longevity. Second, changes in mRNA expression levels and DNA sequences can now be monitored as cells progress along a defined evolutionary path. Such experiments may yield valuable insight into the molecular design of complex control circuits and their adaptation during evolution. The combination of *in silico* and experimental biology introduced here may make a new series of biological designs attainable.

Constraint-based computational models use an optimization-based procedure to predict cellular states. It is assumed that this optimal state, within the governing constraints, is found by altering the numerical values of the kinetic and regulatory constants through a 'trial-and-error' process. This feature of constraint-based models is a significant departure from other types of mathematical models of cell function, where these parameters are treated as being time-invariant. Thus constraint-based models directly account for the fundamental nature of adaptive evolution. The adaptive evolutionary path itself cannot be predicted; however, the final outcome can be. □

Methods

Strains and media

The *E. coli* K-12 MG1655 annotated genome sequence and the biochemical literature were used to construct the *in silico* *E. coli* strain^{1,3,12}. We simulated the metabolic capabilities as previously described with the objective of maximizing growth^{12,14,21}. Both growth and maintenance requirements were imposed on the *in silico* model^{12–14}. The growth experiments were done in M9 minimal medium with the addition of the carbon source. The growth rate was varied by changing the concentration of the carbon source (ranging from 0.25 to 3 g l⁻¹) and the temperature (ranging from 29 °C to 37 °C). *E. coli* MG1655 (ATCC #47076) was used for all of the experiments.

Batch cultures were set up at two different volume scales. One-litre (large) cultures were performed in 1.5-l Erlenmeyer flasks sparging with air. The large-volume batch cultures were used to continuously monitor the oxygen uptake rate (OUR) online with an off-gas analyser. Small (100–250 ml) cultures were grown in 500–1,000 ml Erlenmeyer flasks. For the small-scale cultures the OUR was monitored online polarographically and by measuring the mass transfer coefficient for oxygen (k_{La}) (see below). The temperature was controlled by using a circulating water bath (Haake). We measured and analysed data during exponential growth. The biomass and the concentration of the substrate (malate, glucose, glycerol) in the media were monitored throughout the experiment.

Analytical procedures

Cellular growth rate was monitored by measuring the absorbance (A , or optical density) at 600 nm and 420 nm and by cell counts (Coulter Electronics). The doubling time was calculated from the growth rate: $t_d = \ln(2)/\mu$. Absorbance to cellular dry weight correlations were determined by two different measurements: (1) spun-down cells were dried at 75 °C to a constant weight; and (2) 25–50 ml samples (taken throughout the culture) were filtered, washed and dried to a constant weight. The concentration of metabolites in the culture media was determined by high-performance liquid chromatography (HPLC) (Rainin Instruments). An aminex HPX-87H ion exchange carbohydrate–organic acid column (Bio-Rad Laboratories) (65 °C) was used with degassed 5 mM sulphuric acid as the mobile phase and ultraviolet detection. Glucose and glycerol were monitored by enzymatic assay (Sigma). The dissolved oxygen in the culture was monitored with a polarographic dissolved oxygen probe (Cole-Parmer Instruments). Oxygen consumption was measured in three different ways: (1) passing the effluent gas through a Servomex oxygen analyser (model 1140C) (Servomex); (2) calculated from the dissolved oxygen reading and k_{La} measurements; and (3) in a respirometer chamber in a separate 70-ml flask. All three methods used for measuring the OUR gave similar and reproducible results.

Adaptive evolution

Cultures in prolonged exponential growth were started from individual colonies and were grown in micro-carrier spinner flasks at 250 ml within a temperature-controlled incubator. Serial transfers were made during the exponential phase of growth at mid-log-phase ($A_{600\text{ nm}} = 0.55$) using an adjusted inoculum volume based on the growth rate of the culture. On a daily basis, the growth rate, time of inoculation, $A_{600\text{ nm}}$ of the culture, and any visual changes in the composition of the culture were recorded. Samples of cultures were stored on a daily basis. The culture was tested weekly for pH and phenotyped (by plating) for any signs of coexistence with a distinct population of mutants or foreign contamination. No discernible differences in colony morphology or signs of foreign contamination were observed during these experiments.

Phenotype phase plane analysis

First, the metabolic reconstruction was done using biochemistry, genomic and physiological data^{3–5}. Second, mass balance, capacity and thermodynamic constraints were imposed on the network to define a solution space¹³. Third, the best use of the metabolic network for a given objective was identified using linear optimization^{16–18}. Fourth, all optimal solutions as a function of two constraints were presented on the PPP²¹. The PPP has distinct phases. Each phase corresponds to a particular type of an optimal solution, which in turn represents a particular flux distribution through the network. Each phase has certain metabolic characteristics; for instance, phase 2 in the figures presented here was characterized by an acetate overflow. The lines that demarcate the phases were defined by changes in the shadow price structure of the optimal solution. The properties of the PPP have been detailed elsewhere²¹. In this study, the optimal utilization of the metabolic network was predicted *a priori*, based on a previously developed *in silico* model^{12,14}.

Received 12 December 2001; accepted 2 September 2002; doi:10.1038/nature01149.

- Blattner, F. R. *et al.* The complete genome sequence of *Escherichia coli* K-12. *Science* **277**, 1453–1474 (1997).
- Drell, D. The Department of Energy Microbial Cell Project: a 180° paradigm shift for biology. *OMICS* **6**, 3–9 (2002).
- Covert, M. W. *et al.* Metabolic modeling of microbial strains *in silico*. *Trends Biochem. Sci.* **26**, 179–186 (2001).
- Selkov, E., Maltsev, N., Olsen, G. J., Overbeek, R. & Whitman, W. B. A reconstruction of the metabolism of *Methanococcus jannaschii* from sequence data. *Gene* **197**, GC11–GC26 (1997).
- Overbeek, R. *et al.* WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res.* **28**, 123–125 (2000).
- Karp, P. D. *et al.* The EcoCyc database. *Nucleic Acids Res.* **30**, 56–58 (2002).
- Gombert, A. K. & Nielsen, J. Mathematical modelling of metabolism. *Curr. Opin. Biotechnol.* **11**, 180–186 (2000).
- Tomita, M. *et al.* E-CELL: software environment for whole-cell simulation. *Bioinformatics* **15**, 72–84 (1999).
- Fell, D. *Understanding the Control of Metabolism* (Portland, London, 1996).
- Schuster, S., Fell, D. A. & Dandekar, T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnol.* **18**, 326–332 (2000).
- Schilling, C. H. *et al.* Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.* **184**, 4582–4593 (2002).
- Edwards, J. S. & Palsson, B. O. The *Escherichia coli* MG 1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl Acad. Sci. USA* **97**, 5528–5533 (2000).
- Varma, A. & Palsson, B. O. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl. Environ. Microbiol.* **60**, 3724–3731 (1994).
- Edwards, J. S., Ibarra, R. U. & Palsson, B. O. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotechnol.* **19**, 125–130 (2001).
- Palsson, B. O. The challenges of *in silico* biology. *Nature Biotechnol.* **18**, 1147–1150 (2000).
- Edwards, J. S., Ramakrishna, R., Schilling, C. H. & Palsson, B. O. *Metabolic Engineering* (eds Lee, S. Y. & Papoutsakis, E. T.) (Marcel Dekker, New York, 1999).
- Bonarius, H. P. J., Schmid, G. & Tramper, J. Flux analysis of underdetermined metabolic networks: The quest for the missing constraints. *Trends Biotechnol.* **15**, 308–314 (1997).
- Varma, A. & Palsson, B. O. Metabolic flux balancing: basic concepts, scientific and practical use. *Bio/Technology* **12**, 994–998 (1994).
- Wiechert, W. Modeling and simulation: tools for metabolic engineering. *J. Biotechnol.* **94**, 37–63 (2002).
- Schilling, C. H., Edwards, J. S., Letscher, D. & Palsson, B. O. Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnol. Bioeng.* **71**, 286–306 (2000).
- Edwards, J. S., Ramakrishna, R. & Palsson, B. O. Characterizing the metabolic phenotype: a phenotype phase plane analysis. *Biotechnol. Bioeng.* **77**, 27–36 (2002).
- Weikert, C., Sauer, U. & Bailey, J. E. Use of a glycerol-limited, long-term chemostat for isolation of *Escherichia coli* mutants with improved physiological properties. *Microbiology* **143**, 1567–1574 (1997).

Acknowledgements We thank L. Ramos, J. DiTonno, S. Fong, J. Marciniak, N. Short and H. Bialy for technical assistance and for reviewing the manuscript. We acknowledge funding support from the National Institutes for Health and the National Science Foundation, and the Department of Energy Office of Biological and Environmental Research.

Competing interests statement The authors declare competing financial interests: details accompany the paper on *Nature's* website (► <http://www.nature.com/nature>).

Correspondence and requests for materials should be addressed to B.Ø.P. (e-mail: palsson@ucsd.edu).

Metabolic network structure determines key aspects of functionality and regulation

Jörg Stelling*, Steffen Klamt*, Katja Bettenbrock*, Stefan Schuster† & Ernst Dieter Gilles*

* Max Planck Institute for Dynamics of Complex Technical Systems, D-39106 Magdeburg, Germany

† Max Delbrück Center for Molecular Medicine, D-13092 Berlin, Germany

The relationship between structure, function and regulation in complex cellular networks is a still largely open question^{1–3}. Systems biology aims to explain this relationship by combining experimental and theoretical approaches⁴. Current theories have various strengths and shortcomings in providing an integrated, predictive description of cellular networks. Specifically, dynamic mathematical modelling of large-scale networks meets difficulties because the necessary mechanistic detail and kinetic parameters are rarely available. In contrast, structure-oriented analyses only require network topology, which is well known in many cases. Previous approaches of this type focus on network robustness⁵ or metabolic phenotype^{2,6}, but do not give predictions on cellular regulation. Here, we devise a theoretical method for simultaneously predicting key aspects of network functionality, robustness and gene regulation from network structure alone. This is achieved by determining and analysing the non-decomposable pathways able to operate coherently at steady state (elementary flux modes). We use the example of *Escherichia coli* central metabolism to illustrate the method.

Elementary-mode analysis establishes a link between structural analysis and metabolic flux analysis (MFA). Elementary flux modes can be defined as the smallest sub-networks enabling the metabolic system to operate in steady state⁷. For example, in a hypothetical network (Fig. 1), five elementary modes exist, which cannot further be decomposed. By linear combination of e_1 to e_5 all thermodynamically and stoichiometrically feasible stationary flux distributions can be obtained. In each elementary mode, the enzymes are weighted by the relative fluxes they carry. Up to the non-negative scaling factors for each mode, the set of elementary modes is unique for a given network structure⁸. Hence, it enables us to investigate the space of all physiological states that are meaningful for the cell in the long-term perspective. Flux balance analysis (FBA), in contrast, uses linear programming to obtain a single (not necessarily unique, see Fig. 1) solution to an optimization problem, which is in most cases maximal growth per substrate uptake². Accordingly, FBA focuses on a specific behaviour. It can thus not cope with cellular regulation without additional constraints.

Elementary-mode analysis has mainly been applied to biochemical networks of moderate complexity^{1,7–9}. To explore the utility of the approach for a system of realistic complexity, we chose the central metabolism of the bacterium *Escherichia coli* as an example. In analogy to other network analyses^{2,10}, central carbon metabolism was modelled in (partially extended) detail, whereas in the anabolic part of the model, combining predominantly linear pathways into single assembly reactions served to reduce model complexity. The growth rate is approximated by the production rate of macromolecular cellular constituents such as DNA and protein. We model growth as one reaction converting a fixed ratio of precursors into biomass. Altogether, the network contains 89 substances and 110 reactions, of which 68 reactions can be attributed to single gene products or to multi-enzyme complexes cooperating in a single reaction (see Supplementary Information).

We determined elementary flux modes for the utilization of

representative substrates feeding into different parts of metabolism (Table 1). The total number of elementary modes for given conditions is here used as a quantitative measure of the degrees of freedom¹¹, that is, of flexibility. Glucose, for example, can be used in approximately 45 times more different ways than acetate, corresponding to biological intuition. Flux mode number thus directly relates network structure to function. An empty set implies that no steady-state flux distribution fulfilling the specifications exists, hence predicting an inviable phenotype. For instance, anaerobic use of any of the four substrates except glucose is impossible without additional terminal electron acceptors.

In particular, we analyse the ability to grow or not to grow of mutants carrying deletions in single genes. For this purpose, the number of flux modes for a mutant Δi using substrate S_k is determined by (additionally) selecting for those flux modes that do not require gene i . We denote by $N(\mu, \Delta i)$ the number of flux modes showing a positive growth rate μ for this mutant. The relative number of flux modes for mutant strains (Supplementary Information) allows a correct prediction of the experimentally determined growth phenotype in the overwhelming majority of cases (Fig. 2a). Most situations with an empty (non-empty) set of flux modes correspond to inviable (viable) mutants. The only two false negatives are phosphoglucose-isomerase (*pgi*) mutants, because in the model, growth depends on glucose-6-phosphate production, whereas *in vivo* this precursor is substitutable¹². Erroneous positive predictions may be caused by insufficient pathway capacities (kinetic constraints) *in vivo*. A statistically significant classification of growth behaviour ($P < 10^{-5}$) results from the analysis. Altogether 90% of the predictions (81 out of 90 cases) were correct, which justifies usage of the relative number of elementary modes as a reliable indicator of network function.

As fault-tolerance is an essential feature of living systems, we investigated the structure–function relationship with respect to network robustness. We define robustness as insensitivity of network function, that is, the ability to sustain bacterial growth, towards internal disturbances like mutations¹³. The number of elementary modes qualitatively indicates whether a mutant is viable or not, but it does not necessarily describe to what extent a mutation affects growth quantitatively. We therefore additionally calculated the maximal biomass yield $Y^{\max}$ for each combination of mutant and substrate as a quantitative measure of network performance. Central metabolism of *E. coli* behaves in a highly robust manner, because mutants with significantly reduced metabolic flexibility

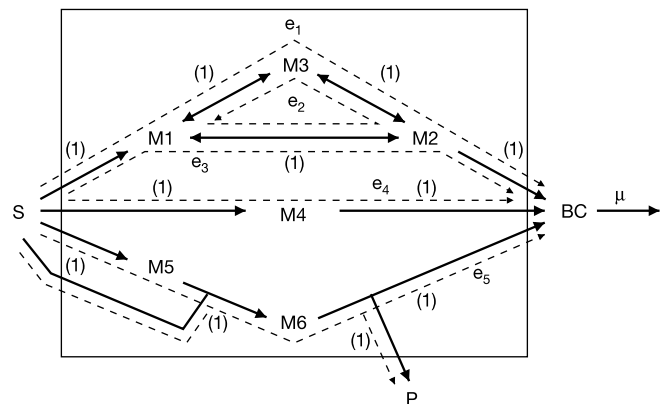


Figure 1 Example network. Reactions (solid arrows, 1:1 stoichiometry for substrates and products) convert substrate S into a biomass component BC and a secreted by-product P via internal metabolites M1–M6. Cellular growth rate μ is approximated by the production of BC. The hypothetical network comprises five elementary flux modes e_i (dashed arrows, relative flux in parentheses). The modes e_1 , e_3 and e_4 give the same BC:S yield of 1:1, while e_5 gives a yield of 1:2.

show a growth yield similar to wild type (Fig. 2b). Robustness relies, at least in part, on pathway redundancy. Analysis of the set of elementary modes in wild type reveals the existence of multiple, alternative pathways with identical biomass yield (Fig. 2c). Only when the number of elementary modes is severely cut down by a mutation is functionality affected. Hence, elementary-mode analysis points to a coexistence of robustness and fragility, as already shown for cellular regulation^{14–16}.

Graph-theoretical methods are widely used to analyse complex networks^{5,17}. In particular, the network diameter D , defined as the average minimal path length between any two nodes (substances), has been shown to be relatively invariant upon random removal of nodes in metabolic networks. This has been suggested to reflect network robustness, as an increasing diameter would indicate network disintegration⁵. However, for the network studied here, a constant diameter does not necessarily imply identical functionality (Fig. 2b). Robustness and fragility, hence, are not predicted by a pure graph-theoretical measure of network topology. In contrast to the network diameter, elementary modes reflect specific characteristics of metabolism such as molar yields. We therefore tackled the question whether the number of elementary modes N directly relates to network robustness. As a measure for robustness, we used the maximal growth yield for each mutant as already shown in Fig. 2b. Counting the number of cases for which $Y^{\max}(\Delta i) > 0$ gave the number of viable single-gene mutants for each substrate regime, that is, a measure of the probability to tolerate random deletion mutations. For different substrate uptake regimes, the organism's resistance to arbitrary gene deletions correlates well ($r^2 = 0.93$) with N for the corresponding wild type (Fig. 2d). The number of elementary modes thus provides an estimate for fault-tolerance.

Finally, we address whether regulation in complex metabolic networks could be predicted by elementary-mode analysis. A direct, quantitative correlation between metabolic fluxes and transcriptome or proteome patterns has not been observed^{10,18,19}. However, the existence of a more indirect link seems likely. We start from the assumption that optimization during biological evolution can be characterized by the two objectives of flexibility—associated with robustness—and efficiency^{11,13}. This is, for example, supported by evidence from the evolution of energy transduction²⁰.

Flexibility means the ability of cellular systems to adapt to a wide range of environmental conditions, that is, to realise a maximal bandwidth of thermodynamically feasible flux distributions, hence of elementary flux modes. Efficiency, as the second objective, could be defined as the fulfilment of cellular demands with an optimal outcome such as maximal cell growth^{2,6}, using a minimum of constitutive elements (such as genes and proteins)¹¹. Since these two criteria impose contradictory challenges, optimal cellular regulation needs to find a trade-off. Our analysis will therefore rely on a parameter characterizing flexibility and efficiency derived from metabolic network structure, for which we introduce the term 'control-effective flux'.

Control-effective fluxes are determined directly from the set of

elementary modes, and do not require optimization. The analysis begins by assigning an efficiency to each elementary mode. These efficiencies relate the mode's output (growth or ATP production) to the investment required to establish the mode, that is, to produce the enzymes. This investment is approximated by the sum of all (absolute) fluxes, because for comparable metabolite concentrations, the flux through an enzymatic reaction scales linearly with enzyme concentration. For the hypothetical network (Fig. 1), elementary mode e_4 would be favoured over e_1 and e_3 involving more enzymatic steps for the same growth yield (see Appendix A, Supplementary Information). Subsequently, we determine control-effective fluxes for a specific reaction as the (normalized) average flux through this reaction in all elementary modes, whereby for each mode the actual flux is weighted by the mode's efficiency (see Methods). In general, control-effective fluxes represent the importance of each reaction for efficient and flexible operation of the entire network. In contrast to FBA, this approach takes network flexibility directly into account because all optimal and sub-optimal modes are considered.

As cellular control on longer timescales is predominantly achieved by genetic regulation, the control-effective fluxes should correlate with messenger RNA levels. Theoretical transcript ratios $\Theta(S_1, S_2)$ for growth on two alternative substrates S_1 and S_2 were therefore calculated as ratios of control-effective fluxes and compared to previously published complementary DNA-microarray data for *E. coli* growing exponentially on glucose, glycerol and acetate^{10,19}. The structure-derived prediction of the differential expression of 50 genes for acetate versus glucose shows good agreement with experiment (Fig. 3a). A test for systematic errors was subsequently performed by comparing the distribution of residuals to the normal distribution resulting from completely

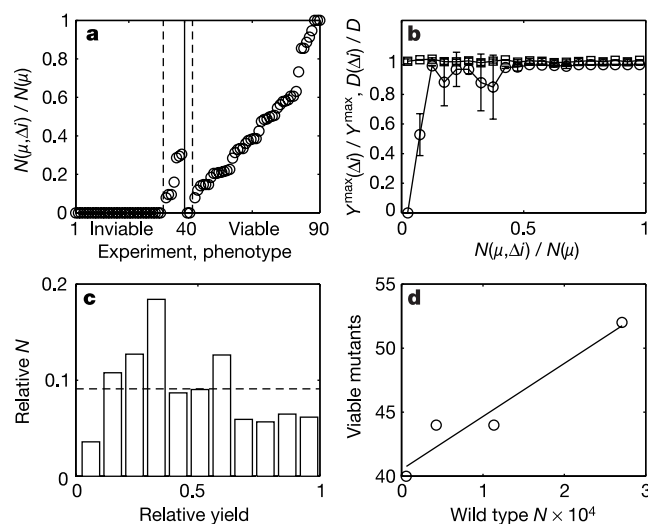


Figure 2 Metabolic network topology and phenotype. **a**, Relative number of elementary modes N enabling deletion mutants in gene i (Δi) of *E. coli* to grow (abbreviated by μ) for 90 different combinations of mutation and carbon source. The solid line separates experimentally determined mutant phenotypes, namely inviability (experiments 1–40) from viability (experiments 41–90). Dashed lines delimit the situations with erroneous predictions. **b**, Dependency of the mutants' maximal growth yield $Y^{\max}(\Delta i)$ (open circles) and the network diameter $D(\Delta i)$ (open squares) on the share of elementary modes operational in the mutants. Data were binned to reduce noise. **c**, Distribution of growth-supporting elementary modes in wild type (rather than in the mutants), that is, share of modes having a specific biomass yield (the dotted line indicates equal distribution). **d**, Effect of arbitrary single-gene deletions on viability for single substrate uptake (open circles) assessed by the mutants' maximal growth yields as in **b**, but considering the numbers of cases in which $Y^{\max}(\Delta i) > 0$, and relating them to the total number of modes for the four substrates in wild type (Table 1).

Table 1 Number and distribution of elementary flux modes.

Selection*		Glucose	Acetate	Glycerol	Succinate	Sum
-	N	27,099	598	11,332	4,249	43,279
Growth only	$N(\mu, \neq ATP)$	73.1%	58.7%	78.6%	76.3%	74.6%
ATP only	$N(\neq \mu, ATP)$	3.2%	5.0%	2.4%	2.4%	3.0%
Growth and ATP	$N(\mu, ATP)$	6.6%	2.0%	5.1%	4.2%	5.9%
No growth/ATP	$N(\neq \mu, \neq ATP)$	17.1%	34.3%	13.9%	17.1%	16.5%
Aerobic growth	$N(\mu, O_2)$	73.1%	60.7%	83.6%	80.5%	76.4%
Anaerobic growth	$N(\mu, \neq O_2)$	6.6%	0.0%	0.0%	0.0%	4.1%

*We denote the number of elementary flux modes simultaneously meeting a set of conditions, $C_1, \dots, C_n$, by $N(C_1, \dots, C_n)$. These conditions include, for example, the situation where cells can grow, which is abbreviated by μ . Excess energy production in the form of ATP (ATP), the substrate metabolized (S_k for the k -th substrate) and oxygen uptake (O_2) are specified accordingly. The operator ' $\neq$ ' indicates that certain fluxes must not occur. The total number of modes includes one futile cycle without substrate uptake.

random deviations²¹. It leads to identification of three presumable outliers (Fig. 3b). Two overestimated transcript ratios are linked to genes involved in acetate metabolism (*pta*, *ackA*) and, hence were expected to be upregulated on acetate. This can be explained by the fact that the gene *acs* codes for an enzyme operating in parallel; for this gene our theoretical prediction is in agreement with experimental observation¹⁹. The transcript ratio of *aspA* encoding for aspartase was underestimated. Elementary-mode analysis suggests that aspartase is required for an effective conversion of excess NADPH generated by the TCA cycle to NADH (Supplementary Information). Residual analysis thus sustained or generated hypotheses amenable to further experimental investigation. Removal of the three outliers from a total of 50 data sets leads to a high correlation between prediction and experiment ($r^2 = 0.60$) with a linear regression close to perfect match. For instance, average expression ratios from independent experimental studies^{10,19} correlate with $r^2 = 0.84$ (not shown).

Additional evidence supporting our approach is provided by three observations. First, predictions based solely on efficiency, namely on the two flux modes with optimal ATP and biomass yield, $\Theta^{\text{opt}}(\text{Ac, Glc})$, displayed a weak correlation (Fig. 3c). Only 28 data points appear because many fluxes are zero and, hence, yield zero or undefined predictions. Such an efficiency analysis corresponds to the approach followed by FBA. Neglecting flexibility may explain why FBA—even when supplied with information on regulatory circuits—only provides qualitative predictions for a subset of genes²². Secondly, in view of the fact that experimental errors are large compared to the effects of changes in the medium, the predictions for other combinations of substrates also agree reasonably well with experiment (Fig. 3d). Finally, prediction quality was poor when for elementary-mode analysis a simultaneous uptake of all substrates was enabled ($N = 507,633$, not shown). The combination of these findings points to a multi-level, hierarchical organ-

isation of metabolic regulation²³. Transcriptional regulation for growth on a specific substrate seems to rely on selection of this substrate regime by the cell, for instance by catabolite repression. According to the substrate regime, gene expression levels are, at an intermediate level of control, adjusted to provide a general set-up for metabolic efficiency and flexibility. At a lower level, short-term regulation of fluxes for a specific situation, such as for one defined substrate concentration, can then be achieved, for example, by allosteric control of metabolic enzymes⁶.

Elementary-mode analysis decomposes complex metabolic networks into simpler units performing a coherent function. The integrative analysis of elementary modes presented here can be used to reconstruct key aspects of cellular behaviour from metabolic network topology, namely to reliably classify mutant phenotypes, to analyse network robustness, and to quantitatively predict functional features of genetic regulation. Including additional knowledge on, for example, newly annotated genes is straightforward⁸. A refined approximation of bacterial growth could serve to improve our method. The concept of extreme pathways²⁴ bears strong similarity with elementary modes^{7,8,11}. For our model of *E. coli* central metabolism, both approaches yield equivalent sets of functional entities and, thus, identical results (not shown). Whereas these approaches characterize the spectrum of different, potential functionalities of the metabolic system, FBA focuses on a single flux distribution. FBA provides similarly good predictions of mutant phenotypes², but it fails whenever network flexibility—for instance, in the analysis of pathway redundancy or in quantitative prediction of gene expression—has to be taken into account. Elementary-mode analysis, in contrast, helps us understand the huge amount of mRNA expression data provided nowadays. Subsequent studies will apply the analysis developed here to organisms other than *E. coli* and further validate it with upcoming transcriptome and mutant data. More generally, we conclude that robustness of metabolic networks is linked to redundancy, and that hierarchical genetic control supports this robustness by finding a trade-off between network efficiency and flexibility. Like recent studies demonstrating the power of combining data and methods from different origins^{25–27}, the systems biology analysis presented here can thus contribute to elucidation of the fundamental design principles of living cells. □

Methods

Network model

For the catabolic part of the model, substrate uptake reactions, glycolysis, pentose phosphate pathway, the TCA cycle with its glyoxylate bypass and the excretion of by-products (acetate, formate, lactate and ethanol) were included. Previous networks² were extended by—among others—the anaplerotic reactions via malic enzyme and pyruvate oxidase as well as by parallel pathways for initial acetate metabolism. The anabolic part of the model covers the conversion of precursors into building blocks like amino acids, to macromolecules and, by assuming a fixed macromolecular composition of the cell, to biomass.

Computation of elementary modes

Elementary flux modes, e_i , as specific vectors r of reaction rates were determined using: (1) the steady-state condition $Sr = 0$ (S is the stoichiometric matrix); and (2) the non-decomposability property that there must not be any steady-state flux vector r that has zero components wherever e_i does and at least one additional zero component^{7,11}. Computation of elementary modes was performed using the FluxAnalyzer²⁸, a program with graphical user interface for the analysis of metabolic networks based on Matlab (Mathworks, Inc.). The software is available upon request from S.K. In particular, a core algorithm described previously⁸ was implemented and optimized by additional pre-processing steps. Unless stated otherwise, flux modes were computed separately for each substrate before being combined.

Robustness analysis

Maximal biomass yield Y^{max} was defined as the optimum of $Y_{i,X/S_k} = e_i^{\text{bi}}/e_i^{\text{S}_k}$. The superscripts to e specify single reaction rates—here, growth and substrate uptake—in elementary flux mode i selected for utilization of substrate S_k . Analysis of network connectivity was performed as described in ref. 5. Results were checked for consistency with topological characteristics of large-scale networks: the network studied here is scale-free, that is, the probability $P(k)$ for a substance to participate in k reactions decays according to $P(k) \approx k^{-\gamma}$, with $\gamma = 1.4$ (Supplementary Information). The diameter of the

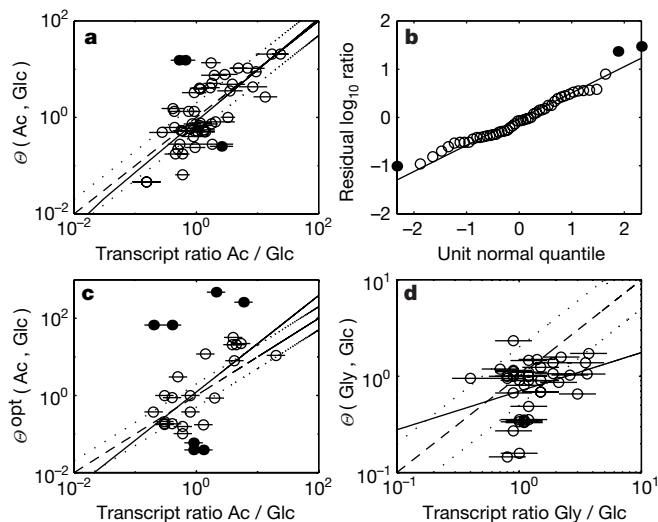


Figure 3 Prediction of gene expression patterns. **a**, Calculated ratios between gene expression levels during exponential growth on acetate and exponential growth on glucose (filled circles indicate outliers) based on all elementary modes versus experimentally determined transcript ratios¹⁹. Lines indicate 95% confidence intervals for experimental data (horizontal lines), linear regression (solid line), perfect match (dashed line) and two-fold deviation (dotted line). **b**, Distribution of residuals, that is, differences between measurement and prediction, compared to a normal distribution of prediction errors. Outliers (filled circles) were identified by their deviation from the linear regression (solid line). **c**, Predicted transcript ratios for acetate versus glucose for which, in contrast to **a**, only the two elementary modes with highest biomass and ATP yield (optimal modes) were considered. **d**, Gene expression ratios for growth on glycerol versus glucose¹⁰ in analogy to **a**.

entire network ($D = 2.9$) corresponds well to network diameters previously reported^{5,29}. We assessed the effect of random mutations on viability by independently determining the number of flux modes, the maximal growth yield, and the network diameter after deletion of each single reaction for growth on glucose, acetate, glycerol or succinate, respectively.

Prediction of gene expression

To calculate control-effective fluxes for each reaction l , we determine the efficiency of any e_i by relating the system's output Ω to the substrate uptake and to the sum of all absolute fluxes. With flux modes normalized by the total substrate uptake, efficiencies $\varepsilon_i(S_k, \Omega)$ for the targets for optimization Ω —growth and ATP generation—are defined as:

$$\varepsilon_i(S_k, \mu) = \frac{e_i^\mu}{\sum_l |e_l^\mu|} \quad \text{and} \quad \varepsilon_i(S_k, \text{ATP}) = \frac{e_i^{\text{ATP}}}{\sum_l |e_l^{\text{ATP}}|} \quad (1)$$

Control-effective fluxes $v_l(S_k)$ are obtained by averaged weighting of the product of reaction-specific fluxes and mode-specific efficiencies over the inventory of elementary modes using the substrate under consideration:

$$v_l(S_k) = \frac{1}{Y_{X/S_k}^{\max}} \cdot \frac{\sum_i \varepsilon_i(S_k, \mu) |e_l^\mu|}{\sum_i \varepsilon_i(S_k, \mu)} + \frac{1}{Y_{A/S_k}^{\max}} \cdot \frac{\sum_i \varepsilon_i(S_k, \text{ATP}) |e_l^{\text{ATP}}|}{\sum_i \varepsilon_i(S_k, \text{ATP})} \quad (2)$$

Here, $Y_{X/S_k}^{\max}$ and $Y_{A/S_k}^{\max}$ denote optimal yields for biomass production and for ATP generation for cellular maintenance, respectively (experimentally determined yield parameters can, however, easily be incorporated into the approach). Subscripts specify elementary-mode number (i) and reaction number (l). Owing to the normalization of modes, the effect of system input (substrate requirement) is implied in equation (2). Theoretical transcript ratios for growth on two alternative substrates S_1 and S_2 are calculated as

$$\Theta_l(S_1, S_2) = \frac{v_l(S_1)}{v_l(S_2)} \quad (3)$$

Received 7 May; accepted 20 September 2002; doi:10.1038/nature01166.

- Dandekar, T., Schuster, S., Snel, B., Huynen, M. & Bork, P. Pathway alignment: Application to the comparative analysis of glycolytic enzymes. *Biochem. J.* **343**, 115–124 (1999).
- Edwards, J. S. & Palsson, B. O. The *Escherichia coli* MG1655 *in silico* metabolic phenotype: its definition, characteristics and capabilities. *Proc. Natl Acad. Sci. USA* **97**, 5528–5533 (2000).
- Cornish-Bowden, A. & Cardenas, M. L. Complex networks of interactions connect genes to phenotypes. *Trends Biochem. Sci.* **26**, 463–465 (2001).
- Kitano, H. Systems Biology: A brief overview. *Science* **295**, 1662–1664 (2002).
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
- Edwards, J. S., Ibarra, R. U. & Palsson, B. O. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotechnol.* **19**, 125–130 (2001).
- Schuster, S., Dandekar, T. & Fell, D. A. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol.* **17**, 53–60 (1999).
- Schuster, S., Fell, D. A. & Dandekar, T. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnol.* **18**, 326–332 (2000).
- Van Dien, S. J. & Lidstrom, M. E. Stoichiometric model for evaluating the metabolic capabilities of the facultative methylotroph *Methylobacterium extorquens* AM1, with application to reconstruction of C₃ and C₄ metabolism. *Biotechnol. Bioeng.* **78**, 296–312 (2002).
- Oh, M.-K. & Liao, J. C. Gene expression profiling by DNA microarrays and metabolic fluxes in *Escherichia coli*. *Biotechnol. Prog.* **16**, 278–286 (2000).
- Heinrich, R. & Schuster, S. *The Regulation of Cellular Systems* (Chapman & Hall, New York, 1996).
- Canonaco, F. *et al.* Metabolic flux response to phosphoglucose isomerase knock-out in *Escherichia coli* and impact of overexpression of the soluble transhydrogenase UdhA. *FEMS Microbiol. Lett.* **204**, 247–252 (2001).
- Barkai, N. & Leibler, S. Robustness in simple biochemical networks. *Nature* **387**, 913–917 (1997).
- von Dassow, G., Meir, E., Munro, E. M. & Odell, G. M. The segment polarity network is a robust developmental module. *Nature* **406**, 188–192 (2000).
- Stelling, J. & Gilles, E. D. in *Proc. 2nd Intl Conf. Syst. Biol.* (eds Yi, T. M., Hucks, M., Morohashi, M. & Kitano, H.) 181–190 (Omnipress, Madison, WI, 2001).
- Csete, M. E. & Doyle, J. C. Reverse engineering of biological complexity. *Science* **295**, 1664–1669 (2002).
- Strogatz, S. H. Exploring complex networks. *Nature* **410**, 268–276 (2001).
- ter Kuile, B. H. & Westerhoff, H. V. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett.* **500**, 169–171 (2001).
- Oh, M.-K., Rohlin, L., Kao, K. C. & Liao, J. C. Global expression profiling of acetate grown *Escherichia coli*. *J. Biol. Chem.* **277**, 13175–13183 (2002).
- Pfeiffer, T., Schuster, S. & Bonhoeffer, S. Cooperation and competition in the evolution of ATP-producing pathways. *Science* **292**, 504–507 (2001).
- Cleveland, W. S. *Visualizing Data* (AT&T Bell Laboratories, Murray Hill, NJ, 1993).
- Covert, M. W. & Palsson, B. O. Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J. Biol. Chem.* **277**, 28058–28064 (2002).
- Stelling, J., Kremling, A., Ginkel, M., Bettenbrock, K. & Gilles, E. D. *Foundations of Systems Biology* (ed. Kitano, H.) 189–212 (MIT Press, Cambridge, MA, 2001).
- Schilling, C. H., Letscher, D. & Palsson, B. O. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.* **203**, 229–248 (2000).
- Marcotte, E. M. *et al.* A combined algorithm for genome-wide prediction of protein function. *Nature* **402**, 83–86 (1999).

- Holter, N. S. *et al.* Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proc. Natl Acad. Sci. USA* **97**, 8409–8414 (2000).
- Ge, H., Liu, Z., Church, G. M. & Vidal, M. Correlation between transcriptome and interactome mapping data from *Saccharomyces cerevisiae*. *Nature Genet.* **29**, 482–486 (2001).
- Klamt, S., Stelling, J., Ginkel, M. & Gilles, E. D. FluxAnalyzer: Exploring structure, pathways, and flux distributions in metabolic networks on interactive flux maps. *Bioinformatics* (in the press).
- Bilke, S. & Peterson, C. Topological properties of citation and metabolic networks. *Phys. Rev. E* **64**, 036106 (2001).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements We thank M. Ginkel for software optimization, J. Liao for providing us with data before publication, U. Sauer, S. Bonhoeffer and A. Cornish-Bowden for critical reading of the manuscript and suggestions. S.S. gratefully acknowledges financial support by the Deutsche Forschungsgemeinschaft.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.S. (e-mail: stelling@mpi-magdeburg.mpg.de).

The heteromeric cyclic nucleotide-gated channel adopts a 3A:1B stoichiometry

Haining Zhong*, Laurie L. Molday† & Robert S. Molday†
King-Wai Yau*‡§

* Departments of Neuroscience and ‡ Ophthalmology, and § Howard Hughes Medical Institute, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA

† Department of Biochemistry and Molecular Biology, and Department of Ophthalmology, University of British Columbia, Vancouver, British Columbia, Canada V6T 1Z3

Cyclic nucleotide-gated (CNG) channels are crucial for visual and olfactory transductions^{1–4}. These channels are tetramers and in their native forms are composed of A and B subunits⁵, with a stoichiometry thought to be 2A:2B (refs 6, 7). Here we report the identification of a leucine-zipper⁸-homology domain named CLZ (for carboxy-terminal leucine zipper). This domain is present in the distal C terminus of CNG channel A subunits but is absent from B subunits, and mediates an inter-subunit interaction. With cross-linking, non-denaturing gel electrophoresis and analytical centrifugation, this CLZ domain was found to mediate a trimeric interaction. In addition, a mutant cone CNG channel A subunit with its CLZ domain replaced by a generic trimeric leucine zipper produced channels that behaved much like the wild type, but less so if replaced by a dimeric or tetrameric leucine zipper. This A-subunit-only, trimeric interaction suggests that heteromeric CNG channels actually adopt a 3A:1B stoichiometry. Biochemical analysis of the purified bovine rod CNG channel confirmed this conclusion. This revised stoichiometry provides a new foundation for understanding the structure and function of the CNG channel family.

Distinct A subunits (or α -subunits) of vertebrate CNG channels, named CNGA1–4 according to the latest nomenclature⁵, mediate rod phototransduction (A1), cone phototransduction (A3) and olfactory transduction (A2 and A4). Two B subunits (or β -subunits) are known, named CNGB1 and CNGB3 (CNGB2 does not exist⁵), with CNGB1 being part of native rod and olfactory channels and CNGB3 part of cone channels. When expressed heterologously, most A subunits (CNGA1–3) form functional homomeric chan-

nels¹⁻⁴; in contrast, B subunits do not, but they form functional heteromers with A subunits. Homologues of these A and B subunits have been identified in *Drosophila*⁹ and *Caenorhabditis elegans*^{10,11}.

While examining possible inter-domain interactions of the human (h)CNGA3 subunit, we found that its cytoplasmic C terminus, designated A3-C, interacted with itself. Thus, when co-expressed in HEK 293T cells, haemagglutinin (HA)-tagged A3-C could be co-immunoprecipitated with Myc-tagged A3-C by an anti-Myc antibody (Fig. 1a). This homotypic interaction was confirmed by an *in vitro* binding assay in which purified recombinant glutathione *S*-transferase (GST)-tagged and His-tagged A3-C were found to interact directly with each other (Fig. 1b). Furthermore, when 293T cells transfected with Myc-tagged A3-C were metabolically labelled and lysed, a single radioactive protein species with the expected molecular mass was immunoprecipitated by the anti-Myc antibody (Fig. 1c), confirming that no other proteins mediated the homotypic interaction. Finally, similar homotypic interactions were

seen with C-terminal constructs of CNGA1 (bovine) and CNGA2 (rat) (Fig. 1d). The C terminus of CNGB1 (human) showed only very weak interaction (data not shown), and that of CNGB3 (mouse) showed no detectable interaction (Fig. 2b, right gel).

Sequence analysis identified a 22-residue leucine-zipper-like sequence (underlined in Fig. 2a) in the C terminus of hCNGA3, downstream of the cyclic-nucleotide-binding site. Deleting these residues abolished the homotypic interaction of a molecule comprising the last 114 residues of A3-C (Fig. 2b, compare A3C-28/A3C-28 and A3C-32/A3C-32 lanes). Further deletion and truncation analyses (Fig. 2b, see A3C-71/A3C-71; also data not shown) revealed that the full interaction domain covers the 22-amino-acid sequence plus 25 residues immediately downstream. This domain, which we name CLZ, is present in all examined CNG channel A subunits (Fig. 2a) but not in the B subunits. It consists of two leucine-zipper-like coiled-coils connected out of phase by a hydrophilic linker. At least several conserved leucine or hydrophobic residues (marked by asterisks in Fig. 2a) are important: when any of them was mutated into alanine, the homotypic interaction was undetectable in the co-immunoprecipitation assay (data not shown). Finally, the CLZ domain was able to confer homotypic interaction on the C terminus of CNGB3 when inserted after the cyclic nucleotide-binding site (Fig. 2b, right gel).

Despite the importance of the CLZ domain, its disruption only

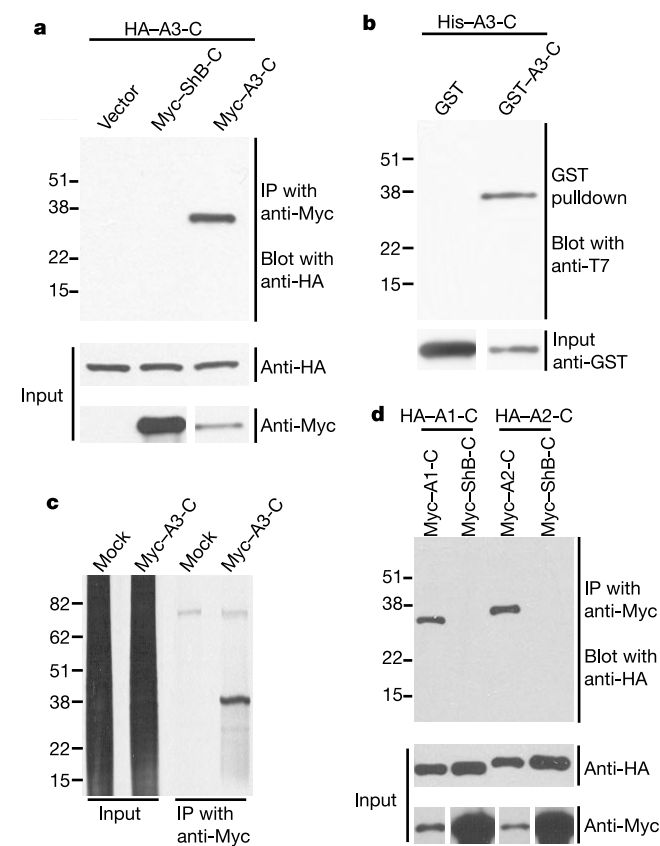


Figure 1 Homotypic interaction of the C terminus of CNG channel A subunit. **a**, Co-immunoprecipitation of epitope-tagged C termini of CNGA3. The constructs were amino-terminally tagged with HA (MGYPYDVPDYADLNGGGGGST) or Myc (MEQKLISEEDLNGGGGGST). The constructs were: A3-C, start of cytoplasmic C terminus (Asn 398) to end of human CNGA3; ShB-C, Asn 482 to end of Shaker B potassium channel (negative control). Inputs before immunoprecipitation (IP) are shown at the bottom and artificially aligned to save space. **b**, Binding *in vitro*. The His-tagged protein was detected with an antibody (69522; Novagen) against a T7 epitope right after the His tag. The inputs of GST (control) and the GST-tagged protein, detected with an anti-GST antibody (27-4577-01; Amersham), before binding *in vitro* are also shown and are artificially aligned (bottom). **c**, Metabolic labelling. 293T cells transfected with Myc-tagged A3-C were metabolically labelled, then immunoprecipitated with the anti-Myc antibody. The pulled-down proteins and inputs before the immunoprecipitation are shown (see Methods). **d**, Co-immunoprecipitation of epitope-tagged C-termini of CNGA1 and CNGA2. The experiment was as described in **a**. The constructs were: A1-C, Asn 393 to end of bovine CNGA1; A2-C, Asn 372 to end of rat CNGA2. Numbers alongside the gels are relative molecular mass (M_r) values in kD.

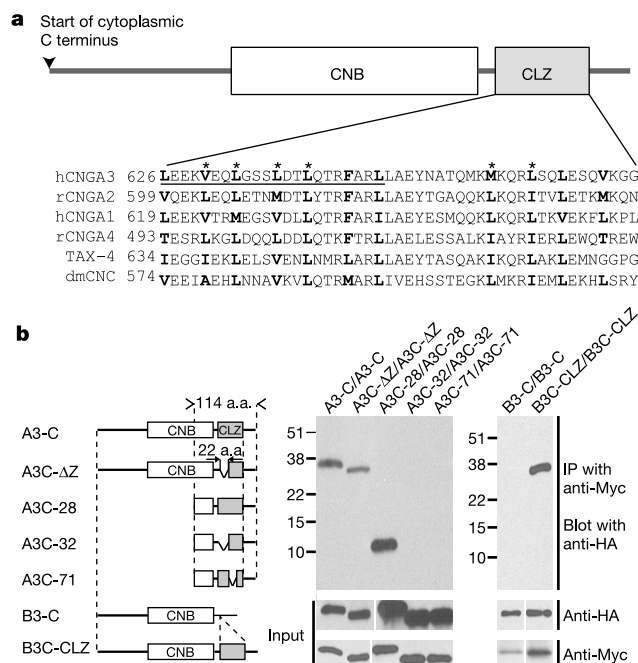


Figure 2 Identification of the CLZ domain. **a**, Cytoplasmic C terminus of CNG channel A subunit, showing the cyclic-nucleotide-binding domain (CNB) and the CLZ domain. Sequence alignments of the CLZ domains of vertebrate and invertebrate CNG channel A subunits are shown underneath. dmCNC and Tax-4 are those of *Drosophila* and *C. elegans*, respectively. Conserved hydrophobic residues are in bold type. The 22-residue leucine zipper first identified by sequence analysis is underlined. Asterisks label hydrophobic residues found to be essential for the CLZ interaction of hCNGA3 in the co-immunoprecipitation assay. Prefixes: h, human; r, rat. **b**, Homotypic interactions of various C-terminal constructs of hCNGA3 and mouse (m)CNGB3. In all lanes, the indicated front construct is HA-tagged and after construct is Myc-tagged. The inputs are arbitrarily aligned. The constructs are: A3C-ΔZ, A3-C with Leu 626 to Leu 647 deleted; A3C-28, Lys 581 to end of hCNGA3; A3C-32, A3C-28 with Leu 626 to Leu 647 deleted; A3C-71, A3C-28 with Leu 648 to Arg 661 deleted; B3-C, Gln 432 to end of mCNGB3; B3C-CLZ, B3-C with Ala 655 to Phe 658 replaced by Lys 624 to Gly 672 of hCNGA3 plus two extra residues (Val-Glu) for facilitating insertion by restriction enzymes. a.a., amino acids. Numbers alongside the gels are M_r values in kD.

weakened, but did not abolish, the interaction of the complete A3C, in that the construct A3C-ΔZ (Fig. 2b, see A3C-ΔZ/A3C-ΔZ lane) still showed a mild interaction. Further experiments suggested that a region upstream of the CLZ domain also interacts homotypically, but the CLZ-mediated interaction is much stronger (data not shown). From here on, we shall focus on the CLZ domain.

In a gel-filtration experiment, a C-terminal fragment (A3C-28; see Fig. 2b) of hCNGA3 having the CLZ domain as the only interacting motif was found to be fractionated into a peak with an apparent molecular mass (estimated from its Stoke's radius) 3.5-fold that of the monomer peak (Fig. 3a; see Methods). Because the heterotetrameric CNG channel contains at least one B subunit, which does not contain the CLZ domain, this result suggests that the fragment probably forms a trimer rather than a tetramer. Indeed, a similar experiment with a slightly larger C-terminal fragment of the channel (T565-End) gave a value of 3.1 (data not shown).

Because the gel-filtration result could have been confounded by molecular shape, we did a cross-linking experiment on A3C-28 with dithiobis-succinimidyl propionate (DSP), a cross-linker cleavable by reducing reagents. Partial cross-linking produced three bands in SDS-PAGE under non-reducing conditions, but only one band in the presence of dithiothreitol (Fig. 3b). These bands presumably represented the monomer, dimer and trimer, respectively. More complete cross-linking shifted the product mostly to the highest (trimer) band. This cross-linking was CLZ-dependent because, when this domain was disrupted (A3C-32 in Fig. 3b), only the lowest (monomer) band was observed. Similar results were obtained for the corresponding C-terminal constructs of hCNGA1 and rat CNGA2 (Fig. 3c), confirming generality among A subunits.

To be definitive, a native-gel approach was used. A3C-28 migrated on a native gel primarily as multimers, as supported by the observation that a point mutant (A3C-87: L633A; Fig. 3d, upper gel) with a disrupted homotypic interaction migrated much more rapidly than the wild type. We constructed two other mutants of A3C-28, retaining the homotypic interaction but with extra negative (A3C-88: K596E, R603E) or positive (A3C-89: D613K, E615K) charges. As expected, A3C-88 had a much higher mobility than A3C-89 in native-gel electrophoresis (Fig. 3d, upper gel). When co-expressed, these two mutants gave four bands (Fig. 3d, upper gel), consistent with the combinatorial possibilities, P_3 , P_2Q , PQ_2 and Q_3 , from two proteins P and Q that form trimers. The four bands became particularly evident when the transfected complementary DNA ratio for A3C-88 and A3C-89 was manipulated (Fig. 3d, lower gel).

The final evidence came from analytical centrifugation, a method of determining the molecular mass independently of shape. A 48-residue peptide (Lys 624–Gly 671) corresponding to the CLZ domain was synthesized and analysed by equilibrium sedimentation. The resulting data were best fitted if the peptide formed a trimer, rather than a dimer or tetramer, in solution (Fig. 3e).

To examine whether the CLZ domain forms a trimer in the context of the complete channel protein, we replaced this domain in hCNGA3 with well-characterized leucine zippers¹² that form dimers, trimers and tetramers, respectively. All mutants readily formed homomeric channels when expressed alone (Fig. 4a). However, the mutants harbouring dimeric and tetrameric leucine zippers consistently differed, although moderately, from the wild-type channel in the $K_{1/2}$ value of the dose–response relation (Fig. 4b) and in the saturated cAMP-activated/saturated cGMP-activated current ratio¹³ (Fig. 4c). In contrast, the mutant harbouring a trimeric leucine zipper mimicked the wild type fairly well in both properties. Furthermore, like the wild type, the trimeric mutant gave comparable currents in the absence and presence of the B subunit, but not the dimeric and tetrameric mutants (Fig. 4d). Thus, the CLZ domain was best emulated by a generic trimeric leucine zipper in functional CNG channels.

The fact that the trimeric CLZ domain is present only in the A subunits but not the B subunits seems to leave only one natural stoichiometry, 3A:1B, for the hetero-tetrameric CNG channel. To examine this hypothesis directly, we purified the native CNG

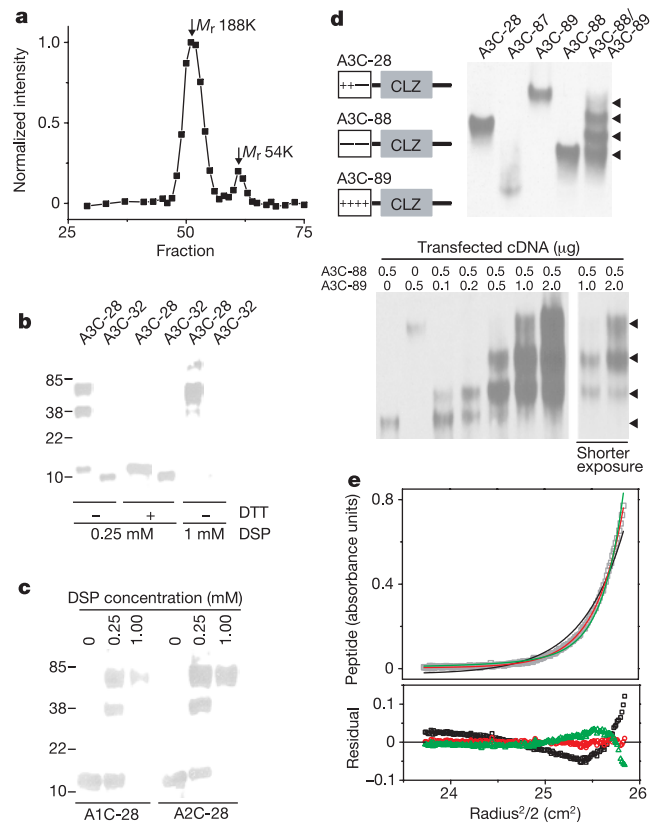


Figure 3 The CLZ domain forms a trimer. **a**, Gel-filtration experiment with N-terminally HA-tagged A3C-28 (see Fig. 2b). See Methods for details. The M_r values of the peaks were estimated from their Stokes radii based on a standard curve (not shown). **b**, Cross-linking experiment on N-terminally HA-tagged A3C-28 or A3C-32 (CLZ-disrupted A3C-28, see Fig. 2b) with the cross-linker DSP. The faint >100K band in the gels seems to be non-specific. DTT, dithiothreitol. **c**, Cross-linking experiments on HA-tagged A1C-28 (Lys 574 to end of human CNGA1) and A2C-28 (Lys 555 to end of rat CNGA2). **d**, Native-gel electrophoresis. All constructs were N-terminally HA-tagged and detected with an anti-HA antibody. A3C-87 is A3C-28 with mutation L633 A, which disrupted the CLZ-mediated interaction without changing the charge of the construct. The position of A3C-87 indicates the monomer position of A3C-28. The two A3C-28 mutants, A3C-88 (K596E and R603E) and A3C-89 (D613K and E615K), had different mobilities in native-gel electrophoresis. When these two constructs are co-expressed, three bands are expected for a dimeric interaction (P_2 , PQ and Q_2) and four bands for a trimer (P_3 , P_2Q , PQ_2 and Q_3). Four bands were seen (indicated by arrowheads) when A3C-88 and A3C-89 were co-expressed. In the bottom panel, a fixed amount of A3C-88 was co-transfected with increasing amounts of A3C-89. Four bands can be clearly seen (arrowheads). **e**, Equilibrium analytical centrifugation. A 48-residue peptide corresponding to the CLZ domain (Lys 624 to Gly 671 of hCNGA3) was commercially synthesized and purified to more than 95% purity (Synpep, Dublin, CA). This peptide was acetylated at the N terminus and amidated at the C terminus, with a calculated M_r of 5,582. The peptide was subjected to analysis by equilibrium analytical centrifugation. A total of nine data sets were collected (0.5, 1.0 and 2.0 mg ml⁻¹, each spun at 40,000, 45,000 and 52,000 r.p.m.). A representative data set (2.0 mg ml⁻¹ centrifuged at 40,000 r.p.m.; open grey squares) and its fitting curves are shown in the upper panel, with residuals (difference between data and fittings) in the lower panel. The data are best fitted by an ideal monomer–trimer model (red) with a M_r of 5,468 (confidence interval 5,375–5,559) and an association constant of 1.43×10^{10} M⁻² (confidence interval $(3.68\text{--}71.7) \times 10^9$ M⁻²). A similar fit with a monomer–dimer model gave an unrealistic M_r in excess of 7,000. With M_r fixed at 5,582 (the calculated value), neither the monomer–dimer (black) nor the monomer–tetramer models (green) fitted well.

channel from bovine rod photoreceptor membranes by affinity chromatography and determined the relative amounts of A and B subunits in the channel complex. The native rod channel consists of a truncated A subunit (CNGA1) and a large B subunit (CNGB1) that migrate on SDS gels as polypeptides of apparent molecular masses 68,000 and 240,000 (M_r 68K and 240K), respectively^{14,15} (Fig. 5a). We quantified the relative tryptophan contents in these bands from the fluorescence arising from ultraviolet-induced trichloroethanol-modified residues¹⁶ (Fig. 5b). The ratio of tryptophan signals for the A:B subunits was 1.5 ± 0.4 (mean $\pm$ s.d.) ($n = 10$), in close agreement with the predicted ratio of 1.3 for a 3A:1B stoichiometry but not with the ratio of 0.4 for a 2A:2B stoichiometry (the truncated A subunit has 9 tryptophan residues; the B subunit has 21). We also determined the relative cysteine contents of the A and B subunits labelled by the fluorescent Oregon Green maleimide (Fig. 5c). The observed modified cysteine ratio of 1.8 ± 0.5 ($n = 3$) for A:B bands was also closer to the predicted ratio of 1.2 for a 3A:1B stoichiometry than the ratio of 0.4 for a 2A:2B stoichiometry (the A subunit has 6 cysteine residues; the B subunit has 15). The higher than expected value observed in our experiments might arise from the limited accessibility of some cysteine residues in the GARP (glutamic-acid-rich protein) part¹⁵ of the B subunit to the bulky Oregon Green maleimide compound. Finally, we examined whether the channel subunits were degraded during purification. Only intact A and B subunits were detected on

western blots labelled with a variety of monoclonal antibodies against different epitopes on these subunits (see Fig. 5d for an example). The absence of degraded fragments indicates that no significant proteolysis of the channel had occurred. Together, our data support a stoichiometry of 3A:1B for the native rod CNG channel.

Our results refute the current belief that the CNG channel adopts a 2A:2B composition, which was largely supported by two studies on the rod channel^{6,7}. One study⁶ relied on a Ni^{2+} potentiation of the opening of homomeric bovine (b)CNGA1 channels by coordinating corresponding histidine residues (His 420) of two adjacent subunits¹⁷; the corresponding residue in bCNGB1 is asparagine (Asn 546). These authors found that Ni^{2+} potentiated the heteromeric channel whether formed by wild-type bCNGA1 and bCNGB1 or by reciprocally mutated bCNGA1 (H420Q) and bCNGB1 (N546H), concluding that there was an AABB stoichiometry/ordering. However, their implicit assumption was that Ni^{2+} acted only on the histidine residue in question, whereas others⁷ have reported that a mild potentiation by Ni^{2+} persisted in the heteromeric channel formed by wild-type bCNGB1 and mutant bCNGA1 (H420Q). The second study⁷ arrived at an ABAB stoichiometry/ordering based on expressions and co-expressions of the tandem dimers A–A, A–B, B–A and B–B. This conclusion assumed that only the leading subunit, not the trailing subunit, of a tandem dimer could incorporate singly into a functional channel. Otherwise, the data would be consistent with a 3A:1B stoichiometry.

In summary, our results suggest that the preferred stoichiometry of heteromeric CNG channels is 3A:1B, although it is impossible to rule out a small percentage of native CNG channels with a different stoichiometry. The olfactory CNG channel is composed of two A subunit species (CNGA2 and CNGA4) and one B subunit species (CNGB1b)^{5,18–21}. Because these A subunits both contain the CLZ domain, the native channel is expected to contain one B subunit together with two A2 and one A4, or one A2 and two A4, subunits (previous work on the stoichiometry of channels formed by co-expressed A2 and A4 is unfortunately not informative because the B subunit was absent²²). Our findings also provide perhaps the first example of a breakdown of symmetry in tetrameric channels. Previous studies based on homomeric CNG channels have suggested that the four CNG subunits in a channel complex are organized into two functional dimers^{23,24}. A three-dimensional structure of purified rod CNG channel at 35 Å resolution²⁵ also suggests that the intracellular domains are arranged as a pair of

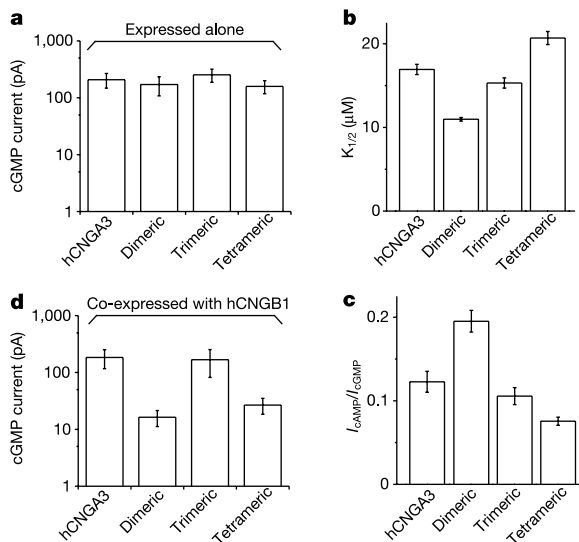


Figure 4 Expression and functional properties of hCNGA3 and its CLZ-substituted mutants. Leu 626 to Val 669 of hCNGA3 was replaced, respectively, by well-characterized dimeric (LEDKVEELLISKNYHLENEVARLKLVGGERI), trimeric (IEDKIEILSKQYHIENELARIKLLIGERI) and tetrameric (IEDKLEEILSKLYHIENELARIKLLIGERI) leucine zippers. All transfections and subsequent recordings were performed side by side. **a**, Saturated cGMP-activated currents of homomeric channels formed by hCNGA3 and the three CLZ-substituted mutants, respectively. The values (means $\pm$ s.e.m.) are 207 ± 60 pA (hCNGA3, $n = 12$), 171 ± 63 pA (dimeric, $n = 12$), 253 ± 66 pA (trimeric, $n = 12$) and 159 ± 41 pA (tetrameric, $n = 13$). **b**, $K_{1/2}$ values of the cGMP dose-response relations for the wild-type and mutant channels. The values (means $\pm$ s.e.m.) are 16.9 ± 0.6 μ M (hCNGA3, $n = 8$), 11.0 ± 0.2 μ M (dimeric, $n = 9$), 15.3 ± 0.6 μ M (trimeric, $n = 9$) and 20.7 ± 0.8 μ M (tetrameric, $n = 9$). **c**, Saturated cAMP/cGMP current ratios of the wild-type and mutant channels. The values (means $\pm$ s.e.m.) are 0.122 ± 0.012 (hCNGA3, $n = 9$), 0.195 ± 0.013 (dimeric, $n = 9$), 0.106 ± 0.010 (trimeric, $n = 9$) and 0.076 ± 0.005 (tetrameric, $n = 9$). **d**, Saturated cGMP current produced by the co-expression of wild-type or mutant hCNGA3 with hCNGB1. The values (means $\pm$ s.e.m.) are 184 ± 67 pA (hCNGA3, $n = 13$), 16.3 ± 5.1 pA (dimeric, $n = 14$), 168 ± 86 pA (trimeric, $n = 14$) and 26.8 ± 8.4 pA (tetrameric, $n = 14$).

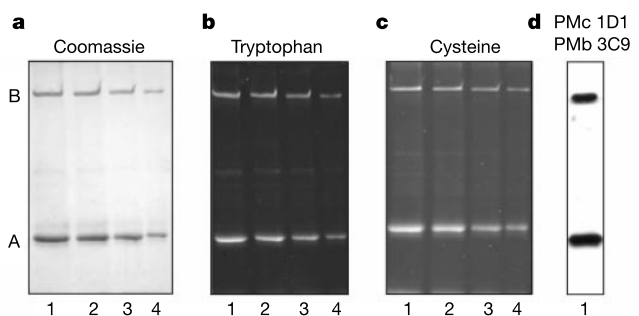


Figure 5 Analysis of the A and B subunits of the native rod CNG channel. The CNG channel was purified from detergent-solubilized bovine rod outer segments on a PMc 1D1–Sepharose immunoaffinity column and analysed by SDS–PAGE and western blotting. **a**, Coomassie Blue-stained gel. **b**, Fluorescence from tryptophan residues modified by an ultraviolet-induced reaction of trichloroethanol and the channel. **c**, Fluorescence from Oregon Green maleimide-modified cysteine residues of the channel. **d**, Western blot labelled with a mixture of monoclonal antibodies, PMb 3C9 and PMc 1D1, against the B ($M_r \sim 240$ K) and A subunits ($M_r \sim 68$ K), respectively. Lanes 1–4 contained about 4, 3, 2 and 1 μ g of protein, respectively.

dimers. Incorporating these findings, the linker between the cyclic-nucleotide-binding site and the CLZ domain might be flexible enough that, despite its asymmetric composition, the other parts of the CNG channel can still configure and function as a 'dimer of dimers'.

Note added in proof: After submission of this paper, we found that ref. 31 also proposed a 3A:1B stoichiometry, although no experimental data were provided. □

Methods

Molecular biology

All of the cDNAs expressed in HEK 293T cells were subcloned into a pRK5 vector. GST fusion proteins were subcloned into pGEX-4T2 (Pharmacia), and His-tagged fusion proteins into pET-28c (Novagene). All mutants were made by standard methods.

Co-immunoprecipitation

The experiments were performed in IP buffer (1 × PBS, 5 mM EDTA, 5 mM EGTA and 1 mM β-mercaptoethanol). At 2 days after transfection using calcium phosphate, the HEK 293T cells were lysed in 1% Triton X-100 and centrifuged at 100,000g for 10 min. The soluble fraction was incubated with a monoclonal anti-Myc antibody (9E10; Roche) and Protein G-Sepharose (Amersham Pharmacia) for more than 2 h. The beads were washed once with 1% Triton, twice with 1% Triton plus 500 mM NaCl, and twice with IP buffer. The precipitate was eluted with Laemmli buffer and subjected to western analysis with the anti-Myc antibody and an anti-HA antibody (2013819; Roche).

In vitro binding

GST and His-tag fusion proteins were prepared in accordance with commercial protocols. Three μg of each protein was added to a final volume of 0.5 ml IP buffer plus 1% Triton. After overnight rocking at 4 °C, glutathione-Sepharose beads (Pharmacia) were added to the reactions and were rocked for 1.5 h. The beads were then washed and eluted similarly to those for the co-immunoprecipitation experiment.

Metabolic labelling

At 2 days after transfection, 293T cells were starved for sulphur for 25 min in a cysteine-free methionine-free medium. A mixture of [³⁵S]cysteine and [³⁵S]methionine (NEG-772; NEN; 1 mCi per 60-mm dish) was added to the medium and growth was allowed for 3 h before the cells were lysed and subjected to immunoprecipitation. The precipitate was separated by SDS-PAGE. The gel was dried before autoradiographic exposure.

Gel filtration

Transfected 293T cells were lysed by freezing and thawing twice. The lysate was rocked at 4 °C for 30 min and centrifuged for 11 min at 100,000g. A fast-protein-liquid-chromatography (FPLC) controller (LCC-500; Pharmacia) was used to fractionate the lysate on a Superdex-200 gel-filtration column (Amersham-Pharmacia) at a flow rate of 0.5 ml min⁻¹. Fractions were collected at half-minute intervals and subjected to western analysis with the anti-HA antibody. The signals on the film were scanned and quantified.

Cross-linking

Transfected 293T cells were lysed as described for gel-filtration experiments. A cross-linking reagent (DSP; Pierce) cleavable by reducing reagents was added to the lysate at different concentrations from a 50-mM stock in dimethyl sulphoxide (DMSO). The reaction was stopped after 30 min by Laemmli buffer with or without reducing reagent (40 mM dithiothreitol). The result was analysed by western blotting with the anti-HA antibody.

Native gel

A continuous gel (10% polyacrylamide) was prepared with the Bio-Rad mini-gel setup. The gel and electrode buffers were 0.1 M Tris-phosphate titrated to pH 7.3 with Tris base. Protein samples were prepared as in gel-filtration experiments, but with added 10% glycerol and 0.01% bromophenol blue (final concentrations) to facilitate loading. The gel was run at a constant 100 V for 7.5 h, then soaked for more than 10 min in buffer used for standard SDS-PAGE (0.1% SDS, 25 mM Tris and 190 mM glycine) to denature the proteins. The gel was subjected to conventional protein transfer and western blotting.

Analytical centrifugation

A peptide was dissolved and dialysed in 1 × PBS, 1 mM EDTA, and samples (0.5, 1 or 2 mg ml⁻¹) were loaded into double-sector centrifugation cells with 12-mm (125-μl) or 3-mm (32-μl) EPON centrepieces and centrifuged at 4 °C at 40,000, 45,000 and 52,000 r.p.m., respectively (Beckman Optima XL-I). Optical absorbance was measured after equilibrium. The data were analysed by global nonlinear least-squares fits with the NONLIN program²⁶. *M_r* and partial specific volume were calculated with Sedenterp version 1.06 (ref. 27).

Electrophysiology

Inside-out excised-patch recordings were made as described²⁸.

Determination of ratio of A:B subunits

The native rod CNG channel was purified from hypotonically lysed bovine rod outer segments by a calmodulin-Sepharose or PMc 1D1 anti-channel monoclonal-antibody-

Sepharose column as previously described²⁹. The samples were run on 7% SDS-PAGE for Coomassie Blue staining and fluorescence detection of modified tryptophan after the ultraviolet-induced reaction of tryptophan with 5% trichloroethanol as described¹⁶. Relative cysteine contents of the A and B subunits were determined by labelling the purified rod CNG channel with 2 mg ml⁻¹ Oregon Green maleimide (Molecular Probes) for 1 hr at 4 °C in the presence of 1% SDS and 20 μg Tris[2-carboxyethyl]phosphine hydrochloride (Pierce) prior to SDS-gel electrophoresis. Oregon Green maleimide labelling of the channel was completely inhibited by previous treatment with 10 mM *N*-ethylmaleimide. The modified tryptophan fluorescence was quantified with a Fluorchem 8000 Digital Imaging System (Alpha Innotech Corp., San Leandro, CA). The fluorescence from the Oregon Green maleimide-labelled channel subunits was quantified with a Typhoon 8600 variable mode imager. In each case, the fluorescence signal was proportional to the amount of protein applied to the gel. The western blots shown in Fig. 5d were labelled with a mixture of the PMc 1D1 and PMb 3C9 monoclonal antibodies against the A and B subunits as described previously^{14,15}. In separate experiments, antibodies against other epitopes were also used but are not shown (PMc 6E7, PMc 2G11, Garp 1H5 and PMs 5E11)^{14,15,30}.

Received 15 August; accepted 1 October 2002; doi:10.1038/nature01201.

1. Finn, J. T., Grunwald, M. E. & Yau, K. W. Cyclic nucleotide-gated ion channels: an extended family with diverse functions. *Annu. Rev. Physiol.* **58**, 395–426 (1996).
2. Zagotta, W. N. & Siegelbaum, S. A. Structure and function of cyclic nucleotide-gated channels. *Annu. Rev. Neurosci.* **19**, 235–263 (1996).
3. Biel, M., Zong, X., Ludwig, A., Sautter, A. & Hofmann, F. Structure and function of cyclic nucleotide-gated channels. *Rev. Physiol. Biochem. Pharmacol.* **135**, 151–171 (1999).
4. Kaupp, U. B. Family of cyclic nucleotide gated ion channels. *Curr. Opin. Neurobiol.* **5**, 434–442 (1995).
5. Bradley, J., Frings, S., Yau, K. W. & Reed, R. Nomenclature for ion channel subunits. *Science* **294**, 2095–2096 (2001).
6. Shammatt, I. M. & Gordon, S. E. Stoichiometry and arrangement of subunits in rod cyclic nucleotide-gated channels. *Neuron* **23**, 809–819 (1999).
7. He, Y., Ruiz, M. & Karpen, J. W. Constraining the subunit order of rod cyclic nucleotide-gated channels reveals a diagonal arrangement of like subunits. *Proc. Natl Acad. Sci. USA* **97**, 895–900 (2000).
8. Alber, T. Structure of the leucine zipper. *Curr. Opin. Genet. Dev.* **2**, 205–210 (1992).
9. Baumann, A., Frings, S., Godde, M., Seifert, R. & Kaupp, U. B. Primary structure and functional expression of a *Drosophila* cyclic nucleotide-gated channel present in eyes and antennae. *EMBO J.* **13**, 5040–5050 (1994).
10. Coburn, C. M. & Bargmann, C. I. A putative cyclic nucleotide-gated channel is required for sensory development and function in *C. elegans*. *Neuron* **17**, 695–706 (1996).
11. Komatsu, H., Mori, L., Rhee, J. S., Akaike, N. & Ohshima, Y. Mutations in a cyclic nucleotide-gated channel lead to abnormal thermosensation and chemosensation in *C. elegans*. *Neuron* **17**, 707–718 (1996).
12. Harbury, P. B., Zhang, T., Kim, P. S. & Alber, T. A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* **262**, 1401–1407 (1993).
13. Gordon, S. E., Oakley, J. C., Varnum, M. D. & Zagotta, W. N. Altered ligand specificity by protonation in the ligand binding domain of cyclic nucleotide-gated channels. *Biochemistry* **35**, 3994–4001 (1996).
14. Molday, R. S. *et al.* The cGMP-gated channel of the rod photoreceptor cell characterization and orientation of the amino terminus. *J. Biol. Chem.* **266**, 21917–21922 (1991).
15. Korschen, H. G. *et al.* A 240 kDa protein represents the complete beta subunit of the cyclic nucleotide-gated channel from rod photoreceptor. *Neuron* **15**, 627–636 (1995).
16. Kazmin, D., Edwards, R. A., Turner, R. J., Larson, E. & Starkey, J. Visualization of proteins in acrylamide gels using ultraviolet illumination. *Anal. Biochem.* **301**, 91–96 (2002).
17. Gordon, S. E. & Zagotta, W. N. Subunit interactions in coordination of Ni²⁺ in cyclic nucleotide-gated channels. *Proc. Natl Acad. Sci. USA* **92**, 10222–10226 (1995).
18. Bradley, J., Li, J., Davidson, N., Lester, H. A. & Zinn, K. Heteromeric olfactory cyclic nucleotide-gated channels: a subunit that confers increased sensitivity to cAMP. *Proc. Natl Acad. Sci. USA* **91**, 8890–8894 (1994).
19. Liman, E. R. & Buck, L. B. A second subunit of the olfactory cyclic nucleotide-gated channel confers high sensitivity to cAMP. *Neuron* **13**, 611–621 (1994).
20. Sautter, A., Zong, X., Hofmann, F. & Biel, M. An isoform of the rod photoreceptor cyclic nucleotide-gated channel beta subunit expressed in olfactory neurons. *Proc. Natl Acad. Sci. USA* **95**, 4696–4701 (1998).
21. Bonigk, W. *et al.* The native rat olfactory cyclic nucleotide-gated channel is composed of three distinct subunits. *J. Neurosci.* **19**, 5332–5347 (1999).
22. Shapiro, M. S. & Zagotta, W. N. Stoichiometry and arrangement of heteromeric olfactory cyclic nucleotide-gated ion channels. *Proc. Natl Acad. Sci. USA* **95**, 14546–14551 (1998).
23. Root, M. J. & MacKinnon, R. Two identical noninteracting sites in an ion channel revealed by proton transfer. *Science* **265**, 1852–1856 (1994).
24. Liu, D. T., Tibbs, G. R., Paoletti, P. & Siegelbaum, S. A. Constraining ligand-binding site stoichiometry suggests that a cyclic nucleotide-gated channel is composed of two functional dimers. *Neuron* **21**, 235–248 (1998).
25. Higgins, M. K., Weitz, D., Warne, T., Schertler, G. F. & Kaupp, U. B. Molecular architecture of a retinal cGMP-gated channel: the arrangement of the cytoplasmic domains. *EMBO J.* **21**, 2087–2094 (2002).
26. Johnson, M. L., Correia, J. J., Yphantis, D. A. & Halvorson, H. R. Analysis of data from the analytical ultracentrifuge by nonlinear least-squares techniques. *Biophys. J.* **36**, 575–588 (1981).
27. Laue, T. M., Shah, B. D., Ridgeway, T. M. & Pelletier, S. L. *Analytical Ultracentrifugation in Biochemistry and Polymer Science* (eds Harding, S. E., Rowe, A. J. & Horton, J. C.) 90–125 (Royal Society of Chemistry, Cambridge, UK, 1992).
28. Chen, T. Y. & Yau, K. W. Direct modulation by Ca²⁺-calmodulin of cyclic nucleotide-activated channel of rat olfactory receptor neurons. *Nature* **368**, 545–548 (1994).
29. Molday, R. S. & Molday, L. L. Purification, characterization, and reconstitution of cyclic nucleotide-gated channels. *Methods Enzymol.* **294**, 246–260 (1999).

30. Poetsch, A., Molday, L. L. & Molday, R. S. The cGMP-gated channel and related glutamic acid-rich proteins interact with peripherin-2 at the rim region of rod photoreceptor disc membranes. *J. Biol. Chem.* **276**, 48009–48016 (2001).
31. Kaupp, U. B. & Seifert, R. Cyclic nucleotide-gated ion channels. *Physiol. Rev.* **82**, 769–824 (2002).

Acknowledgements We thank R. Haganir, P. Kim, D. Leahy, M. Li, J. Nathans, F. Rupp and D. Yue for advice and suggestions; members of the Yau laboratory for comments on the manuscript; M. Biel for the mCNGB3 cDNA; P. Kim for providing us with peptide samples corresponding to dimeric, trimeric and tetrameric leucine zippers; and M. Rodgers for helping us in the analytical centrifugation experiment. This work was supported by grants from the US National Eye Institute to K.-W.Y. and R.S.M.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.S.M. (e-mail: molday@unixg.ubc.ca). or K.-W.Y. (e-mail: kyau1@jhem.jhmi.edu).

Establishment of transcriptional competence in early and late S phase

Jianmin Zhang^{*†}, Feng Xu^{*†‡}, Tamar Hashimshony^{*}, Ilana Keshet^{*} & Howard Cedar^{*}

^{*} Department of Cellular Biochemistry and Human Genetics, Hebrew University, Ein Kerem, Jerusalem, Israel

[†] These authors contributed equally to this work

In animal cells, the process of DNA replication takes place in a programmed manner, with each gene region designated to replicate at a fixed time slot in S phase. Housekeeping genes undergo replication in the first half of S phase in all cell types, whereas the replication of many tissue specific genes is developmentally controlled, being late in most tissues but early in the tissue of expression¹. Here we employ nuclear DNA injection as an experimental system to test whether this phenomenon is due to differences in the ability to set up transcriptional competence during S phase^{2,3}. Our results show that, regardless of sequence, exogenous genes are a better template for transcription when injected into nuclei of cells in early as opposed to late S phase, and this expression state, once initiated, is preserved after cell division. DNA injected in late S phase is apparently repressed because it is packaged into chromatin containing deacetylated histones, and the same is true for late replicating chromosomal DNA. These findings suggest a mechanistic connection between replication timing and gene expression that might help to explain how epigenetic states can be maintained *in vivo*.

Non-synchronized Rat-1 cells were plated onto a marked gridded disk at a density permitting the identification of each individual cell by its position on the plate. Cells were randomly injected with an S16–LacZ reporter plasmid and each one was followed by microscopy at 2-hour intervals to determine its time of division and thereby ascertain its cell cycle position at the time of injection (Fig. 1a). At 20 h after injection, cells were fixed and stained for LacZ activity and the number of positive cells was graphed as a function of cell cycle stage at the time of injection.

S16–LacZ gene sequences injected into early-S-phase cells were more efficiently transcribed (~10-fold) than when injected into late-S-phase cells (Fig. 1b). Regardless of the stage of the cell cycle at which the injection took place, activity was always determined after 20 h, so that in each case the cells went through a full cycle and had

the same amount of time to accumulate product. Another expression vector made up of the EF1- α housekeeping promoter linked to a green-fluorescent-protein (GFP) reporter showed a similar difference between injection during early and late S phase (data not shown), and the same was true for tissue specific promoters such as α -actin or myogenin (Fig. 1b), indicating that this is a general, sequence-independent phenomenon. Expression competence, once established, seems to remain fixed even though the cell continues to go through its normal division cycle. Thus, late-injected cells retain a low transcription rate even 48 h after injection, whereas the early-injected substrate maintains its high level of expression (Table 1).

The fact that templates injected into late-S-phase cells remain relatively inactive, even after many hours in the cell, strongly suggested that transcriptional competence is established and fixed soon after injection. To prove this directly, we designed a new injection protocol (Fig. 2a) that allowed us to isolate early-injected and late-injected cells and assay transcription directly by polymerase chain reaction with reverse transcription (RT–PCR) (Fig. 2b). After 3 h, early-S-injected S16- β -globin was already more transcriptionally efficient (6.7-fold) than a template injected during late S

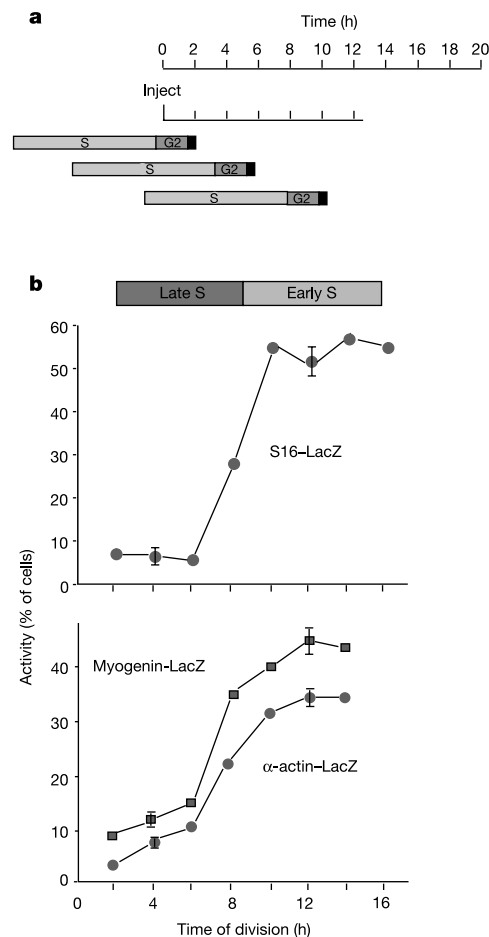


Figure 1 Transcriptional competence as a function of S phase. **a**, Cells on a Celloclate grid were injected (long vertical bar) at zero time and followed every 2 h (short bars) to determine whether mitosis (black) had occurred. Mitosis at 10 h (bottom cell) indicates that injection took place in early S. **b**, Rat-1 cells were injected with S16–LacZ ($0.75 \text{ ng } \mu\text{l}^{-1}$; 700 cells), myogenin–LacZ ($1.25 \text{ ng } \mu\text{l}^{-1}$; 850 cells) or α -actin–LacZ ($0.8 \text{ ng } \mu\text{l}^{-1}$; 1,250 cells) and assayed for LacZ-positive cells after 20 h. Error bars indicate the 95% confidence interval for the 2-h, 4-h and 6-h or the 10-h, 12-h and 14-h points as a group. The diagram at the top indicates cells injected at early S (dark grey) or late S (light grey).

[‡] Present address: Toronto General Hospital, 200 Elizabeth Street, MBRC 1-934, Toronto, Canada ONT M5G 2C4.

phase, and similar results were obtained even after 1 h (data not shown). These differences were not simply due to effects of the cell cycle on template stability, because exogenous plasmids were present in equal amounts whether injected early or late in S phase (Fig. 2c).

As a final test of the idea that it is indeed the time of injection that determines transcriptional competence, we designed an experiment to measure the expression of early S and late S templates in a single cell. Whereas cells injected with the S16–LacZ reporter construct in early S were very efficient in initiating transcription (63%), only 19% of this same cell population expressed an E1F- α –GFP plasmid injected in late S (Fig. 2d). Similar results were obtained when the GFP reporter was injected in early S and the S16–LacZ in late S, or even when the protocol was reorganized so that the first injection was directed into cells in late S. These results clearly indicate that DNA initially exposed to late S remains relatively inactive, even though new vectors injected into early S are highly transcribed in the same cells.

We next examined the chromatin structure of injected DNA. Using quantitative PCR analysis we first showed that exogenous templates are present in nucleosomal DNA at about the same level as endogenous sequences (β -actin), in both early-injected and late-injected cells (Fig. 3a).

We then characterized histone acetylation^{4,5} by employing two plasmids that differed from each other by a small deletion (Fig. 3b). One plasmid (I) was injected into cells in early S, and the other (II) was injected in late S (exp. 1). Nucleosomes from the combined cell populations were immunoprecipitated with anti-Ac-H4 and assayed by quantitative PCR, using β -actin and Ig- κ as active

Table 1 Transcriptional competence as a function of S phase

Gene	Assay time (h)	LacZ-positive cells (%)	
		Injection in early S	Injection in late S
S16–LacZ	24	69 $\pm$ 8	22 $\pm$ 8
	48	65 $\pm$ 9	17 $\pm$ 6
Myogenin–LacZ	24	46 $\pm$ 10	14 $\pm$ 6
	48	43 $\pm$ 10	9 $\pm$ 4

Rat-1 cells were injected with S16–LacZ (0.75 ng μ l⁻¹) or myogenin–LacZ (1.25 ng μ l⁻¹) and assayed for LacZ-positive cells after 24 or 48 h; means $\pm$ 95% confidence interval are shown. Cells that divided within the first 6 h after injection were considered late-S recipients, whereas those that divided within the next 6 h were considered early-S recipients.

(enriched) and inactive endogenous gene controls. Early-injected DNA was about 10-fold more enriched for acetylated histones than late-injected DNA (Fig. 3b and d). Similar results were obtained in a reciprocal injection experiment (exp. 2) or when S16–LacZ was used as the template (see legend to Fig. 3). These findings indicate that DNA injected in early and late S adopt different chromatin structures and that this is not dependent on the presence of an active promoter.

These studies indicated that the repression of late-S-injected reporter genes might be due to the presence of deacetylated histones. To test this hypothesis, we injected the S16–LacZ reporter gene into cells in early or late S that had been pretreated with a short pulse of trichostatin A (TSA). This treatment led to a striking increase (2.8-fold) in expression of late-S-injected DNA (Fig. 4a), indicating that histone deacetylation partly accounts for the repression phenomenon. However, once this state is set up in late S it remains stable and even subsequent exposure to TSA several hours

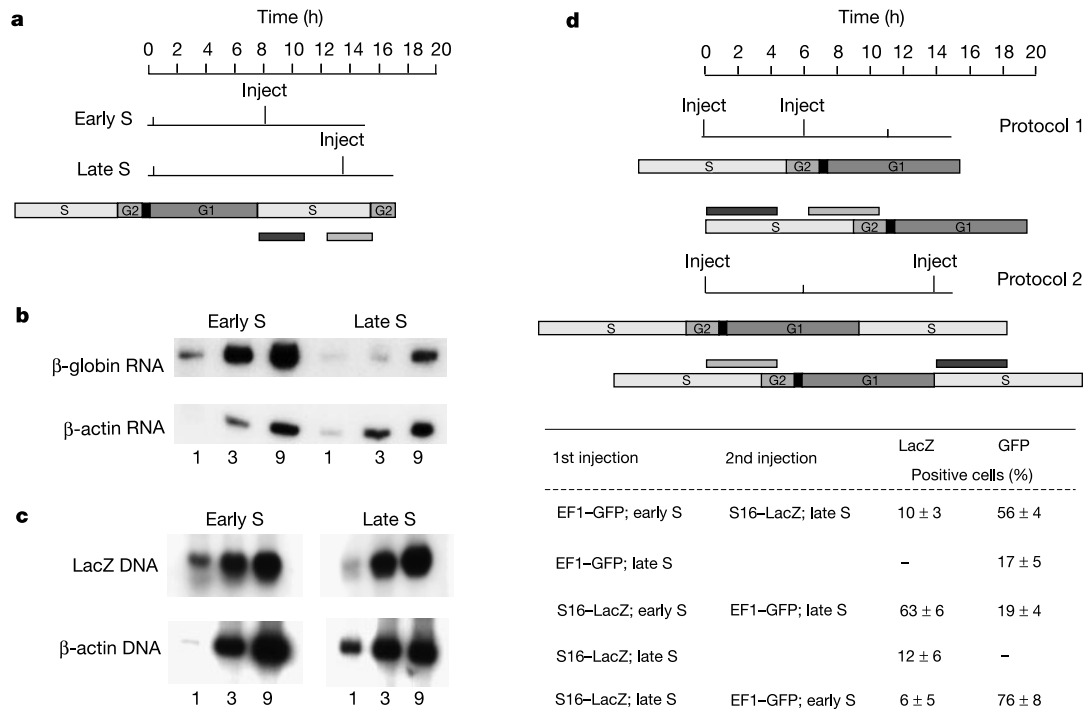


Figure 2 Characterization of injected DNA. **a**, Recently divided cells (within 3 h) were identified (short vertical bar) by their morphology, and their positions recorded. Injection (long bar) was performed after either 8 h (early S cells) or 13 h (late S cells). Early (dark-shaded) and late (light-shaded) injected time intervals are about 3 h long. **b**, Cells were injected with S16– β -globin (2 ng μ l⁻¹) in early S (320 cells) or late S (250 cells) and harvested after 3 h. Quantitative RT–PCR (1, 3 and 9 μ l) was performed after adjusting for injected cell number. **c**, S16–LacZ was injected in late S or early S (equal numbers) and assayed for DNA after 3 h. Similar results were obtained for an S16– β -globin template 10

or 24 h after injection. **d**, Rat-1 cells were injected at zero time (early S) with either S16–LacZ (0.75 ng μ l⁻¹) or EF1- α –GFP (0.4 ng μ l⁻¹). At 6 h, non-divided cells were injected for a second time (late S) and cells that divided within the next 5 h were recorded (protocol 1). The injection time spans for early S (dark-shaded area) and late S (light-shaded area) are shown. In protocol 2, cells were injected at zero time (late S) and those that divided within 6 h were marked and then injected 8 h later (14 h on the time scale) (early S). Results ($\pm$ 95% CI) from 250–380 divided cells for each point are tabulated.

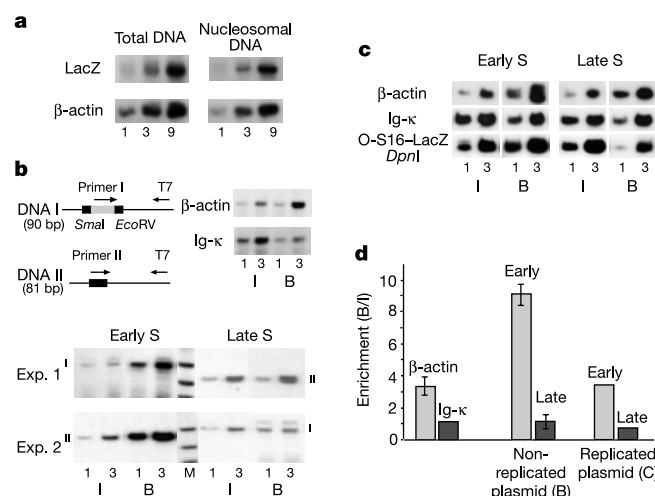


Figure 3 Cell-cycle-dependent histone acetylation. **a**, Nuclei (500) from a tissue culture plate containing 10^5 cells were injected with S16-LacZ, isolated after 24 h and used to prepare nucleosomal DNA for PCR analysis. The ratios of nucleosomal to nuclear DNA for each sequence were 0.9 (β -actin) and 0.8 (S16-LacZ). Almost identical results were obtained for dinucleosomes and trinucleosomes and for S16-LacZ DNA injected (24 h) in early S (0.8) or late S (0.9). **b**, Histone acetylation on newly packaged DNA was assayed after injection (Fig. 2a) of pBluescript II SK (DNA-I) or the same plasmid containing a small deletion near the T7 primer (DNA-II). DNA-I was injected into early-S cells and DNA-II into late-S cells (exp. 1) or the reverse (exp. 2). At 24 h after injection both cell populations were combined and nucleosomes were immunoprecipitated with anti-Ac-H4 (ref. 27). Total input (I) and bound DNA (B) were analysed²⁸ by quantitative PCR (two concentrations). **c**, Histone acetylation was assayed after injection of O-S16-LacZ into E1/E2-expressing Rat-1 cells in early or late S. Replicated molecules were assayed by digestion with *DpnI* before PCR. **d**, Summary of results (means $\pm$ s.e.) in **b** and **c** relative to Ig- κ (set at 1.0).

after injection (post) has no effect. A similar experiment was done by injecting the S16- β -globin reporter and assaying early-injected and late-injected cells by RT-PCR (Fig. 4b). Here, too, TSA treatment overcomes much of the repression observed in late-S-injected cells.

To determine whether differential S-phase nucleosomal packaging also takes place on replicating molecules, we generated a plasmid carrying the bovine papillomavirus (BPV) origin of replication (O-S16-LacZ) and injected it into early or late S-phase Rat-1 cells made replication-competent by previous transfection with an E1/E2-expressing episomal vector⁶. Because the injected DNA is of bacterial origin, it is initially digestible by *DpnI* (G^m ATC) but becomes resistant after replication in animal cells. By this criterion, about 30–40% of the plasmid molecules injected in either early or late S-phase have undergone replication within the first 3 h (data not shown). Mononucleosomes from these cells were immunoprecipitated with anti-Ac-H4 and extracted DNA was then digested with *DpnI* to eliminate all except the replicated plasmid molecules. Once again, we observed that early-S DNA is assembled into nucleosomes enriched with Ac-H4, whereas late-S-replicated vectors become associated with deacetylated histones (Fig. 3c and d). Similar results were also observed for Ac-H3 (data not shown).

In the light of these results, we then asked whether S-phase control also has a role in setting up the acetylation pattern of nucleosomes assembled on replicating chromosomal DNA. It has already been demonstrated that DNA within late replication bands is specifically associated with deacetylated histone H4, both in general and on the X chromosome in particular⁷, but we wished to devise a sequence-independent test of this idea. Rat-1 cells were

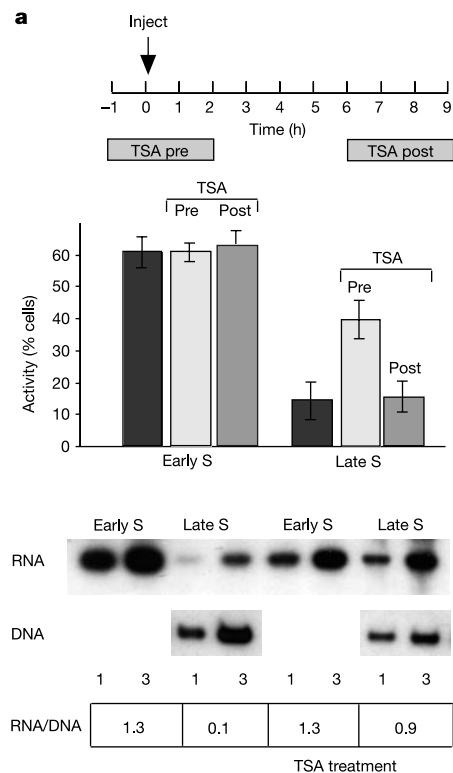


Figure 4 TSA relieves late-S repression. **a**, Rat-1 cells were injected as in Fig. 1 with S16-LacZ (700 cells) or without (800 cells) a 3-h TSA (50 ng ml^{-1}) treatment initiated either 1 h before (pre) or 6 h after (post) injection. Late-S cells were those that divided 2–6 h after injection; early-S injected cells were those dividing 8–12 h after injection. **b**, Cells were injected with S16- β globin as in Fig. 2a and analysed by RT-PCR after normalization to β -actin. TSA was added 1 h before injection. For each experiment, cells injected early and late had similar amounts (by PCR) of DNA (only late is shown).

first stably transfected with a neutral plasmid (pBluescript), and a pool of colonies representing about 200 individual integration events was grown up. We assumed that, in these cells, some insertion sequences replicate in early S, whereas others undergo DNA synthesis in late S, and we wished to determine whether this distribution is correlated with histone acetylation. We therefore performed chromatin immunoprecipitation (ChIP) analysis with anti-Ac-H4 on separate pooled colonies incubated for 1 h in 5-bromodeoxyuridine (BrdU) in either early or late S phase (Fig. 5). Strikingly, early-replicating BrdU-labelled plasmid copies were found to be enriched for acetylated histones, whereas the sequence-identical late-replicating copies in the same cells were depleted. These studies show that, like episomal DNA, chromosomal sequences are specifically packaged into a deacetylated histone structure when replicated in late S.

The original 'window of opportunity' model was based on the concept that cell-cycle variations in transcription factor availability can directly influence the probability of forming stable transcription complexes^{2,3}. Our studies support a chromatin-based promoter-independent model whereby DNA becomes packaged differently in early S as opposed to late S. The difference in competence is evidently caused by cell-cycle-dependent factors that influence chromatin structure⁸ by altering histone modification^{4,9}. One possibility is that, whereas all newly replicated DNA might be initially packaged with Ac-H4 (refs 5, 10), nucleosomes assembled in late S are specifically deacetylated¹¹. This is consistent with experiments showing that the histone deacetylase HDAC2 is localized exclusively to late-S replication complexes^{12,13}.

The process of setting up transcriptional competence most probably represents an integral part of the normal replication

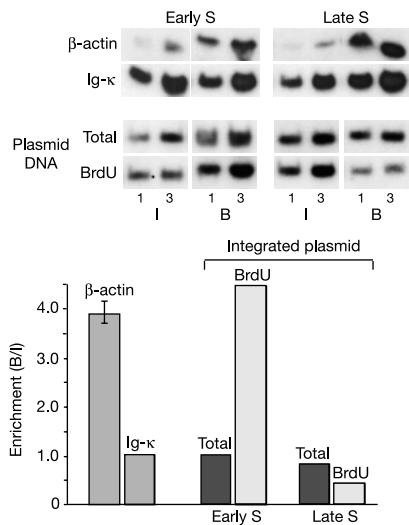


Figure 5 Histone acetylation of early-replicating and late-replicating chromosome DNA. Pools of Rat-1 colonies transfected with the plasmid pBluescript were incubated with BrdU (1 h) during either early or late S after synchronization (see Methods), and cells were harvested 3 h later. ChIP (anti-Ac-H4) was performed on mononucleosomes and enrichment was determined for total plasmid sequences (which should be the same in both cell populations); early or late BrdU-labelled plasmids were isolated by immunoprecipitation of input (I) and bound (B) DNA with anti-BrdU. Eighty per cent of the transfected plasmid sequences are integrated in late-replicating regions. Enrichment (graph) was calculated relative to Ig-κ (set to 1.0) in each cell population.

cycle whereby endogenous chromatin structure is first disrupted and then reassembled in the wake of DNA synthesis^{14,15}. The genomic replication timing profile of the genome itself seems to be set up independently of transcription^{16,17} by means of *cis*-acting elements¹ and *trans*-acting factors^{18–20} that affect the firing of local origins. As a result of this developmentally regulated organization and because the ability to set up transcriptional competence is dependent on S phase, early-replicating genes become assembled into an active structure, whereas late-replicating genes might be packaged into an inactive form during each passage through the cell cycle, thus providing a time-axis-based global mechanism for maintaining epigenetic states in dividing cells. □

Methods

Plasmids

pBluescript-based reporter constructs carrying the GFP or LacZ (β -galactosidase) coding sequence together with a nuclear localization signal were used to insert promoter regions for the mouse ribosomal protein gene S16 (208 base pairs; bp), the rat α -actin gene (800 bp), the mouse myogenin gene (1,100 bp) or the human EF1- α gene (1,207 bp) by PCR or restriction enzyme technology. S16- β globin was made by inserting the S16 promoter next to the human β -globin gene sequence²¹, and O-S16-LacZ by inserting the *EcoRI/BamHI* BPV origin-containing fragment derived from plasmid pKSO²² into S16-LacZ at the *NotI* site. The BPV-derived vector, B45-Neo, which lacks the genes E5, E6 and E7, was transfected into Rat-1 cells and stably maintained in episomal form at about 30 copies per cell⁶.

Microinjections

Rat-1 cells (10^5) were seeded on 5.5-cm tissue culture plates containing gridded Cellocate disks (Eppendorf) 20 h before injection. Under these conditions each disk contains about 150–200 well-spaced cells, each of which can be identified from markings on the disk. Nuclear injection (0.4 – 1.5 ng μl^{-1} DNA) was performed as described²³. Calibration of the system by using radioactive material showed that the average injection volume was 20 fl, containing 3–12 molecules of supercoiled plasmid DNA. For each plasmid we first performed trial injections at various concentrations to determine the amount of DNA that yielded suboptimal reporter gene activity (~ 50 – 60% LacZ- or GFP-positive cells). About 70% of the injected cells remained viable and divided normally in culture.

In experiments employing LacZ reporter genes, cells were fixed, usually 20 h after injection, and then stained with 5-bromo-4-chloro-3-indolyl- β -D-galactoside for 24 h before recording the positive reaction²⁴. GFP-positive cells were detected by fluorescence microscopy 20 h after injection. However, identical results could be obtained by measuring

GFP even 6 h after injection. It should be noted that only vital cells, those that had divided in the course of the experiment, were recorded. Injected DNA lacking an origin sequence remains unreplicated for at least 48 h as shown by restriction analysis with *DpnI* and *MboI* before PCR. The BPV-origin-containing plasmid O-S16-LacZ, in contrast, underwent $\sim 30\%$ replication within 3 h of injection into B45-neo-transfected Rat-1 cells in either early or late S phase.

To programme injection schedules it was necessary to determine the cell cycle properties of these cells. We first determined the length of a full cell cycle by microscopic observation (on 100 cells) of the time from cell division to cell division. The lengths of G2 and S were then measured by incubating a non-synchronous population of cells with 5×10^{-5} M BrdU for 1 h, and then performing immunostaining²⁵ to detect incorporation into cells that divided at 2-h intervals after labelling. With this assay we found that uninjected cells have a cell cycle of 16 h (7 h G1, 7.5 h S, 1.5 h G2). A similar experiment on cells injected immediately after incubation with BrdU indicated that the injection procedure actually extends the cell cycle to 19 h (8 h G1, 9 h S, 2 h G2).

In some experiments cells were treated with Nocodazole ($0.2 \mu\text{M}$) and then synchronized at M/G1 by mitotic shake-off²⁶. More than 80% of these cells then cycled normally and could be injected in either early or late S on separate plates. Monomeric and multimeric nucleosome fractions were prepared and ChIP analysis was performed with anti-Ac-H4 or anti-Ac-H3 (Upstate Biotechnology) to separate input from bound fractions as described²⁷. Because $\sim 1\%$ of the nucleosomes are precipitated by this procedure, PCR of the bound fraction was always compared with a 1:100 dilution of the input DNA. To measure the acetylation state of chromosomal DNA, Rat-1 cells were stably transfected with pBluescript and grown as a pool of ~ 200 individual colonies. Cells were synchronized by mitotic shake-off (above) and then incubated with 10^{-5} M BrdU for 1 h in either early or late S phase. ChIP analysis was performed with anti-Ac-H4 and replicated sequences were specifically detected by immunoprecipitation of BrdU-containing DNA.

PCR analysis

PCR analysis of DNA was performed on LacZ ($5'$ -ACGGCATGGTCCAAATGAAT- $3'$; $5'$ -GACCAGATGATCACACTCGG- $3'$, 127 nucleotides (nt), 36 cycles), endogenous β -actin ($5'$ -CGCCATGGATGACGATATCG- $3'$; $5'$ -CGAAGCCGGCCTTGACATG- $3'$, 68 nt, 21 cycles or 32 cycles for ChIP), β -globin ($5'$ -GCTTCTGACACAACCTGTGTTC- $3'$; $5'$ -CTGAAGTTCTCAGGATCCACG- $3'$, 450 nt, 36 cycles), Ig- κ ($5'$ -AAACCTGCCTGAA GCCGAGC- $3'$; $5'$ -GATTGTGGGAGGAATGAGAA- $3'$, 76 nt, 32 cycles), pBluescript in Fig. 4 (primer I, $5'$ -GGGCTGCAGGAATTCGAT- $3'$, 90 nt; primer II, $5'$ -GGATCCCCCAT CAAGCTTATCG- $3'$, 81 nt; T7, $5'$ -GTAATACGACTCACTATAGGG- $3'$, 40 cycles) and pBluescript in Fig. 5 ($5'$ -CGCTCTAGAACTAGTGGATC- $3'$; $5'$ -TCGAGGTCGACGG TATC- $3'$, 72 nt, 38 cycles). RT-PCR²¹ was performed on β -globin (as above, 352 nt, 36–39 cycles) and endogenous β -actin ($5'$ -CGCCATGGATGACGATATCG- $3'$; $5'$ -CACGTAG GAGTCCTTCTGAC- $3'$, 166 nt, 23 cycles) after optimizing conditions. β -Globin and β -actin RNAs were detected with inter-exon primers. Quantification was performed by scanning autoradiograms or by PhosphorImager analysis and averaging the results for two or three different concentrations. To examine replicated O-S16-Lac Z plasmid specifically, DNA was treated before PCR with *DpnI* ($10 \mu\text{U} \mu\text{l}^{-1}$) for 4 h. Under these conditions, at least 85% of the bacterial DNA is digested (data not shown).

Received 13 March; accepted 5 September 2002; doi:10.1038/nature01150.

- Simon, I. & Cedar, H. In *DNA Replication in Eukaryotic Cells* (ed. DePamphilis, M. L.) 387–408 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1996).
- Gottesfeld, J. & Bloomer, L. S. Assembly of transcriptionally active 5S RNA gene chromatin *in vitro*. *Cell* **28**, 781–791 (1982).
- Brown, D. D. The role of stable complexes that repress and activate eucaryotic genes. *Cell* **37**, 359–365 (1984).
- Roth, S. Y. & Allis, C. D. Histone acetylation and chromatin assembly: a single escort, multiple dances? *Cell* **87**, 5–8 (1996).
- Grunstein, M. Histone acetylation in chromatin structure and transcription. *Nature* **389**, 349–352 (1997).
- Ohe, Y., Zhao, D., Saijo, N. & Podack, E. R. Construction of a novel bovine papillomavirus vector without detectable transforming activity suitable for gene transfer. *Hum. Gene Ther.* **6**, 325–333 (1995).
- Jeppesen, P. & Turner, B. M. The inactive X chromosome in female mammals is distinguished by a lack of histone H4 acetylation, a cytogenetic marker for gene expression. *Cell* **74**, 281–289 (1993).
- Workman, J. L. & Kingston, R. E. Alteration of nucleosome structure as a mechanism of transcriptional regulation. *Annu. Rev. Biochem.* **67**, 545–579 (1998).
- Lee, D. Y., Hayes, J. J., Pruss, D. & Wolffe, A. P. A positive role for histone acetylation in transcription factor access to nucleosomal DNA. *Cell* **72**, 73–84 (1993).
- Sobel, R. E., Cook, R. G., Perry, C. A., Annunzio, A. T. & Allis, C. D. Conservation of deposition-related acetylation sites in newly synthesized histones H3 and H4. *Proc. Natl Acad. Sci. USA* **92**, 1237–1247 (1995).
- Wolffe, A. P. In *DNA Replication in Eukaryotic Cells* (ed. DePamphilis, M. L.) 271–293 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1996).
- Rountree, M. R., Bachman, K. E. & Baylin, S. B. DNMT1 binds HDAC2 and a new co-repressor, DNMT1, to form a complex at replication foci. *Nature Genet.* **25**, 269–277 (2000).
- Allshire, R. & Bickmore, W. Pausing for thought on the boundaries of imprinting. *Cell* **102**, 705–708 (2000).
- Weintraub, H. Assembly of an active chromatin structure during replication. *Nucleic Acids Res.* **7**, 781–792 (1979).
- Solomon, M. J. & Varshavsky, A. A nuclease-hypersensitive region forms de novo after chromosome replication. *Mol. Cell. Biol.* **7**, 3822–3825 (1987).
- Simon, I. *et al.* Developmental regulation of DNA replication timing at the human β globin locus. *EMBO J.* **20**, 6150–6157 (2001).
- Cimbara, D. M. *et al.* Long-distance control of origin choice and replication timing in the human

- beta-globin locus are independent of the locus control region. *Mol. Cell. Biol.* **20**, 5581–5591 (2000).
18. Donaldson, A. D. *et al.* CLB5-dependent activation of late replication origins in *S. cerevisiae*. *Mol. Cell* **2**, 173–182 (1998).
 19. Santocanale, C. & Diffley, J. F. A Mec1- and Rad53-dependent checkpoint controls late-firing origins of DNA replication. *Nature* **395**, 615–618 (1998).
 20. Shirahige, K. *et al.* Regulation of DNA-replication origins during cell-cycle progression. *Nature* **395**, 618–621 (1998).
 21. Siegfried, Z. *et al.* DNA methylation represses transcription *in vivo*. *Nature Genet.* **22**, 203–206 (1999).
 22. Yang, L., Li, R., Mohr, I. J., Clark, R. & Botchan, M. R. Activation of BPV-1 replication *in vitro* by the transcription factor E2. *Nature* **353**, 628–632 (1991).
 23. Graessmann, M. & Graessmann, A. Microinjection of tissue culture cells. *Methods Enzymol.* **101**, 482–492 (1983).
 24. Kutsukake, M., Komatsu, A., Yamamoto, D. & Ishiwa-Chigusa, S. A tyramine receptor gene mutation causes a defective olfactory behavior in *Drosophila melanogaster*. *Gene* **245**, 31–42 (2000).
 25. Simon, I. *et al.* Asynchronous replication of imprinted genes is established in the gametes and maintained during development. *Nature* **401**, 929–932 (1999).
 26. Brandeis, M. & Hunt, T. The proteolysis of mitotic cyclins in mammalian cells persists from the end of mitosis until the onset of S phase. *EMBO J.* **15**, 5280–5289 (1996).
 27. Hebbes, T. R., Clayton, A. L., Thorne, A. W. & Crane-Robinson, C. Core histone hyperacetylation co-maps with generalized DNase I sensitivity in the chicken beta-globin chromosomal domain. *EMBO J.* **13**, 1823–1830 (1994).
 28. Eden, S., Hashimshony, T., Keshet, I., Cedar, H. & Thorne, A. W. DNA methylation models histone acetylation. *Nature* **394**, 842–843 (1998).

Acknowledgements We thank B. Joshua, S. Beyit and B. Giloh for their help in developing the injection technology, and O. Meyuhas, M. Brandeis, E. Bachrach and D. Yaffe for providing vectors used in this study. This research was supported by grants from the N.I.H., the Israel Academy of Sciences and the Israel Cancer Research Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.C. (e-mail: cedar@md2.huji.ac.il)

erratum

Reassessing the evidence for the earliest traces of life

Mark A. van Zuilen, Aivo Lepland & Gustaf Arrhenius

Nature **418**, 627–630 (2002).

On page 628, line 23, of this Letter, the isotope equilibrium fractionation temperature of 500 °C was incorrectly stated as 1,500 °C. □

corrigenda

Recovery of 16S ribosomal RNA gene fragments from ancient halite

Steven A. Fish, Thomas J. Shepherd, Terry J. McGenity & William D. Grant

Nature **417**, 432–436 (2002).

Figure 4b of this Letter included a sequence (Thailand-2 AJ319571), which we now realise is highly likely to be a chimaera between sequences Thailand-6 (AJ319575) and Thailand-7 (AJ319576), also shown in the phylogenetic tree. The chimaeric sequence is one out of twenty-three halite sequences shown in Fig. 4b, and did not occupy a pivotal position in the discussion; its inclusion in the phylogenetic tree therefore does not affect our general conclusions. The chimaeric sequence has now been removed from the databases. We are grateful to E. Willerslev and A. Cooper for bringing this to our attention. □

IRE1 couples endoplasmic reticulum load to secretory capacity by processing the *XBP-1* mRNA

Marcella Calfon, Huiqing Zeng, Fumihiko Urano, Jeffery H. Till, Stevan R. Hubbard, Heather P. Harding, Scott G. Clark & David Ron

Nature **415**, 92–96 (2002).

In this Letter, the commercial polyclonal antiserum used to detect endogenous XBP-1 was sc-7160 (Santa Cruz Biotechnology) and not sc-8015, as erroneously stated. The antibody sc-8015 is a mouse monoclonal and, in our hands, is not useful in detecting the endogenous protein by immunoblotting. □

Stars of the screen

The latest releases for high-throughput analysis.

PrepExpress

Applied Biosystems

www.appliedbiosystems.com

High-throughput compound characterization

Designed for the API 150EX PrepLC/MS system, the Windows-based PrepExpress application for Analyst software will automatically purify hundreds of compounds weekly for combinatorial and medicinal chemistry applications. A prescreening analytical experiment determines crude sample purity, and the software provides automatic submission of impure compounds for preparative purification, as well as automatic purity analysis of the collected fractions. The API 150EX PrepLC/MS system is configured in both analytical and preparative modes, with two Shimadzu LC8-A liquid chromatography pumps and a Gilson 215 autosampler/fraction collector.

Tissue microarrays

InnoGenex

www.innogenex.com

Screening *in situ* gene expression

InnoGenex now produces tissue microarrays for high-throughput screening of *in situ* gene expression for target validation. Tissue microarrays are available from diseased and normal human tissues, mouse normal tissues and rat normal tissues. They are suitable for rapid screening of *in situ* differential expression of genes, identifying novel antibody markers, animal model studies, or for any other application that would use traditional tissue sections. Low-density arrays contain 20–50 elements per slide and high-density arrays have 100–200 elements per slide. First to be available in this new range are tissue arrays with normal and diseased lung, normal and diseased oesophagus, normal and diseased cervix, normal and diseased colon, normal human tissue arrays, and rat and mouse normal visceral organs for whole animal survey.



Eppendorf: Perfect for plasmid purification.

Perfectprep Plasmid 384 Kit

Eppendorf

www.eppendorf.com

Plasmids for high throughput

The Perfectprep Plasmid 384 kit was developed for the high-throughput purification of plasmids using a liquid handling workstation. By using technology that ensures a high surface area, the kit overcomes the problems usually associated with the 384-well format and generates consistent high yields and optimal concentrations. The average yield with the Perfectprep Plasmid 384 kit is 1 µg of high-copy plasmid from only 0.3 ml of bacterial culture. Since the elution volume is only 10 µl, the average concentration is 100 ng per µl. The kit protocol features full 384-well format processing starting with bacteria cultivation (96-well cultivation is an option).

MultiScreen Caco-2 Assay

Millipore

www.millipore.com

Higher throughput ADME screening

Millipore's MultiScreen Caco-2 assay system performs *in vitro* Caco-2 drug transport assays. The system is validated and quality control released with Caco-2 cell lines in 10-day and 21-day cultures. Active drug screening results correlate with 24-well systems and meet FDA values for known compounds.

Optimized to grow and sustain high-integrity Caco-2 cell monolayers, the MultiScreen Caco-2 assay includes the necessary growth plates, feeding tray and transport analysis plate to complete the process from growth to analysis in one plate.

ShortFast HPLC columns

Grace Vydac

www.grace.com

High-throughput analyses in protein/peptide research

This new line of rapid analysis HPLC columns was developed for rapid separation of proteins and peptides. Applications include reaction kinetics, combinatorial library screening and proteomics. The makers say that ShortFast columns allow purification of more material, more quickly than ever before. Compared with analysis using conventional chromatography columns, ShortFast columns are claimed to reduce run times by as much as 85%, and they also require less solvent usage, resulting in disposal cost savings.

SNPCube

Proteodyne

www.proteodyne.com

High-throughput SNP genotyping

The SNPCube automated high-throughput SNP genotyping system completely automates single nucleotide polymorphism (SNP) assays based on the TaqMan series of assays from Applied Biosystems. Included in the range is the SNPCube-2 system, incorporating assay preparation, multiplexed PCR thermal cycling, sequence detection, complete data tracking and data integration with Proteodyne's XML LIMS interface layer. Speeds of up to 90,000 SNP reactions per day can be achieved.

These notes are compiled in the Nature office from information provided by the manufacturers.

nature insight

computational biology



This Insight presented us with a difficult problem, not in its content — a collection of reviews showing how sophisticated mathematical concepts have illuminated and continue to illuminate the principles underlying biology at a genetic, molecular, cellular and even organismal level. The problem was what to call it.

There is considerable interest in this sort of biology at the moment, with well-funded centres springing up at a number of prestigious universities. Most commonly it is referred to as 'systems biology', relating it to systems engineering. But such a term is far too all inclusive, as when biology ceases to concern itself with the 'systems' of organisms it ceases to be biology and becomes instead a subdiscipline of chemistry or physics.

Equally 'Mathematical Biology' or 'Quantitative Biology' didn't fit the bill, as quantitative measurements and their mathematical and statistical manipulation underlie science as a whole. Someone suggested 'Holistic Biology' or even 'Whole-istic Biology', but saner council prevailed.

We also did not want a name that implied that this was a new topic. Physiologists have been looking at the functioning of organisms as a whole for decades, if not centuries. Applying network analysis to cell signalling, metabolism and genetics features heavily in the Insight, but Stuart Kauffman and others were pioneering such approaches in the 1960s. As far back as 1902, Theodor Boveri tested the chromosomal theory of inheritance with probabilistic simulations.

In the end we concluded that the unifying strand that runs through all the work described in this Insight was computation, whether it be the production of sophisticated models against which reality is compared, or the subtle analyses that derive patterns and trends from vast and noisy data sets. There are other themes running through the reviews in this Insight, more than you might expect from the titles alone, but 'Computational Biology' it has become.

Given their promotion and encouragement of this discipline, we are pleased to acknowledge the financial support of NIGMS and NHGRI in producing this Insight. As always, *Nature* carries sole responsibility for all editorial content and peer review.

Christopher Surridge **Senior Editor**

overview:

- 206 Computational systems biology**
H. Kitano

review articles:

- 211 The language of genes**
D. B. Searls

- 218 The structure of the protein universe and genome evolution**
E. V. Koonin, Y. I. Wolf & G. P. Karev

- 224 Engineered gene circuits**
J. Hasty, D. McMillen & J. J. Collins

- 231 Control, exploitation and tolerance of intracellular noise**
C. V. Rao, D. M. Wolf & A. P. Arkin

- 238 Computational approaches to cellular rhythms**
A. Goldbeter

- 246 The community of the self**
T. G. Buchman

Editor, *Nature*: Philip Campbell
Insights Editors: Lesley Anson
Ursula Weiss
Editorial Assistant: Simon Gibson

Art Director: Majo Xeridat
Diagrams: Ann Thomson
Vicky Askew
Suzanne Coleman

Production Editor: Simon Gribbin
Production: Sue Gray
Marketing: Claire Aspinall
Sponsorship: Mark Greene

Computational systems biology

Hiroaki Kitano

Sony Computer Science Laboratories Inc., 3-14-13 Higashi-gotanda, Shinagawa, Tokyo 141-0022, ERATO Kitano Symbiotic Systems Project, Japan Science and Technology Corporation, and The Systems Biology Institute, Suite 6A, M31, 6-31-15 Jingu-mae, Shibuya, Tokyo 150-0001, School of Fundamental Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama, Kanagawa 223-8522, Japan, and Control and Dynamical Systems, California Institute of Technology, Pasadena, California 91125, USA (e-mail: kitano@csl.sony.co.jp)

To understand complex biological systems requires the integration of experimental and computational research — in other words a systems biology approach. Computational biology, through pragmatic modelling and theoretical exploration, provides a powerful foundation from which to address critical scientific questions head-on. The reviews in this Insight cover many different aspects of this energetic field, although all, in one way or another, illuminate the functioning of modular circuits, including their robustness, design and manipulation. Computational systems biology addresses questions fundamental to our understanding of life, yet progress here will lead to practical innovations in medicine, drug discovery and engineering.

It is often said that biological systems, such as cells, are 'complex systems'. A popular notion of complex systems is of very large numbers of simple and identical elements interacting to produce 'complex' behaviours. The reality of biological systems is somewhat different. Here large numbers of functionally diverse, and frequently multifunctional, sets of elements interact selectively and nonlinearly to produce coherent rather than complex behaviours.

Unlike complex systems of simple elements, in which functions emerge from the properties of the networks they form rather than from any specific element, functions in biological systems rely on a combination of the network and the specific elements involved. For example, p53 (a 393-amino-acid protein sometimes called 'the guardian of genome') acts as tumour suppressor because of its position within a network of transcription factors. However, p53 is activated, inhibited and degraded by modifications such as phosphorylation, dephosphorylation and proteolytic degradation, while its targets are selected by the different modification patterns that exist; these are properties that reflect the complexity of the element itself. Neither p53 nor the network functions as a tumour suppressor in isolation. In this way, biological systems might be better characterized as symbiotic systems.

Molecular biology has uncovered a multitude of biological facts, such as genome sequences and protein properties, but this alone is not sufficient for interpreting biological systems. Cells, tissues, organs, organisms and ecological webs are systems of components whose specific interactions have been defined by evolution; thus a system-level understanding should be the prime goal of biology. Although advances in accurate, quantitative experimental approaches will doubtless continue, insights into the functioning of biological systems will not result from purely intuitive assaults. This is because of the intrinsic complexity of biological systems. A combination of experimental and computational approaches is expected to resolve this problem.

A two-pronged attack

Computational biology has two distinct branches: knowledge discovery, or data-mining, which extracts the hidden patterns from huge quantities of experimental data, forming hypotheses as a result; and simulation-based analysis, which tests hypotheses with *in silico* experiments, providing predictions to be tested by *in vitro* and *in vivo* studies.

Knowledge discovery is used extensively within bioinformatics for such tasks as the prediction of exon–intron and protein structure from sequence¹, and the inference of gene regulatory networks from expression profile^{2–4}. These methods typically use predictions based on heuristics, on statistical discriminators that often involve sophisticated approaches (such as hidden Markov models) and on other linguistic-based algorithms (see review in this issue by Searls, pages 211–217).

In contrast, simulation attempts to predict the dynamics of systems so that the validity of the underlying assumptions can be tested. Detailed behaviours of computer-executable models are first compared with experimental observation. Inconsistency at this stage means that the assumptions that represent our knowledge on the system under consideration are at best incomplete. Models that survive initial validation can then be used to make predictions to be tested by experiments, as well as to explore questions that are not amenable to experimental inquiry.

Although traditional bioinformatics has been used widely for genome analysis, simulation-based approaches have received little mainstream attention. This is now changing. Current experimental molecular biology is now producing the high-throughput quantitative data needed to support simulation-based research. Combined with rapid progress of genome and proteome projects, this is convincing increasing numbers of researchers of the importance of a system-level approach⁵. At the same time, substantial advances in software and computational power have enabled the creation and analysis of reasonably realistic yet intricate biological models.

There are still issues to be resolved, but computational modelling and analysis are now able to provide useful biological insights and predictions for well understood targets such as bifurcation analysis of the cell cycle^{6,7}, metabolic analysis^{8,9} or comparative studies of robustness of biological oscillation circuits¹⁰.

It is crucial that individual research groups are able to exchange their models and create commonly accepted repositories and software environments that are available to all. Systems Biology Markup Language (SBML; <http://www.sbml.org/>), CellML (<http://www.cellml.org/>) and the Systems Biology Workbench are examples of efforts that aim to form a *de facto* standard and open software platform for modelling and analysis^{11,12}. These significantly increase the value of the new generation of databases concerned with biological pathways, such as the Kyoto

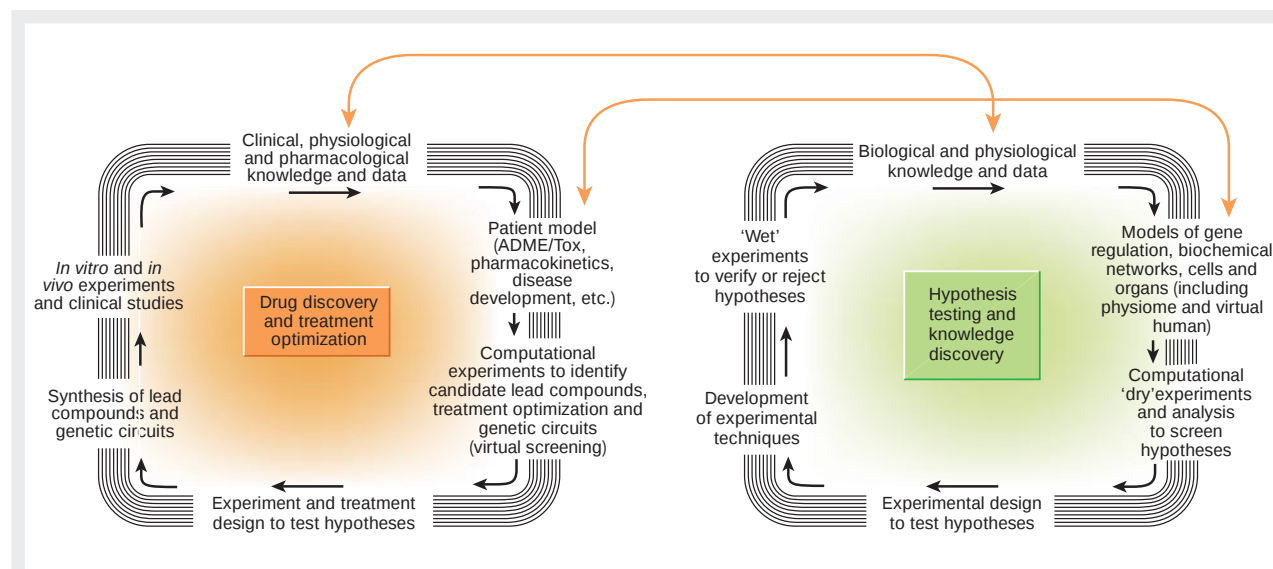


Figure 1 Linkage of a basic systems-biology research cycle with drug discovery and treatment cycles. Systems biology is an integrated process of computational modelling, system analysis, technology development for experiments, and quantitative experiments¹⁸. With sufficient progress in basic systems biology, this cycle can be applied to drug discovery and the development of new treatments. In the future, *in silico* experiments and screening of lead candidates and multiple drug systems, as well as introduced genetic circuits, will have a key role in the 'upstream' processes of the pharmaceutical industry, significantly reducing costs and increasing the success of product and service development.

Encyclopedia of Genes and Genomes (KEGG)¹³, Alliance for Cellular Signaling (AfCS)¹⁴ and Signal Transduction Knowledge Environment (STKE)¹⁵, by enabling them to develop machine-executable models, rather than mere human-readable forms.

Such changes are fuelling a renewed interest in a system-level approach to biology, but we should not forget that this is an area with a long history^{16,17}, rooted as much as anywhere in classical physiology (see review in this issue by Buchman, pages 246–251). However, the close linkage between system-level understanding and molecular-level knowledge was made possible only by the recent progress in genomics and proteomics. The approach attempts to understand biological systems as systems, specifically targeting the identification of their structures and dynamics, and the establishment of methods to control cellular behaviours by external stimuli and to design genetic circuits with desired properties. These aims will be achieved only by combining computation, system analysis, new technologies for comprehensive and quantitative measurements, and high-throughput quantitative experimental data^{18,19}.

Multiple faces of robustness

Among various scientific questions, one issue receiving considerable attention is how robustness is achieved and how it evolves within various aspects of biological systems. Robust systems maintain their state and functions against external and internal perturbations, and robustness is an essential feature of biological systems, having been studied since the earliest attempts at a system-oriented view (for example, Cannon's homeostasis and Wiener's cybernetics¹⁶). Biological systems have been found to be robust at a variety of levels from genetic switches to physiological reactions (see review in this issue by Buchman, pages 246–251).

Robust systems are both relatively insensitive to alterations of their internal parameters and able to adapt to changes in their environment. In highly robust systems, even damage to their very structure produces only minor alterations in their behaviour. Such properties are achieved through feedback, modularity, redundancy and structural stability.

A variety of feedback and feed-forward control is observed throughout biology. For example, integral feedback is central to bacteria chemotaxis^{20–22}. And p53-based cell-cycle arrest displays what is

known in the engineering field as 'bang-bang control', a subtype of feedback control. Damage to DNA is sensed by proteins such as ATM (for ataxia telangiectasia mutated, named after a disease in which this enzyme is mutated) and DNA-dependent protein kinase, which activate the p53 protein. Active p53 then transactivates p21, which results in G1 arrest; this state is released when DNA damage is repaired, thus forming a feedback loop.

Cells themselves provide the most obvious form of biological modularity by physically partitioning off biochemical reactions. However, biochemical networks within cells also form modular compartments isolated by spatial localization²³, anchoring of proteins to plasma membranes and by dynamics.

Cells also provide redundancy, with many autonomous units carrying out identical roles. But redundancy also appears at other levels by having multiple genes that encode similar proteins, or multiple networks with complementary functions. For example, *Per1*, *Per2* and *Per3* genes encode proteins in the circadian oscillator, but knock-out of one or two of these produces no visible phenotype. The *Cln* gene family form redundant pairs for the cell cycle²⁴. The stringent response of *Escherichia coli* activates alternative metabolic dynamics depending upon the availability of lactose and glucose²⁵.

Structurally stable network configurations increase insensitivity to parameter changes, noise and minor mutations. For example, elegant experiments on the archetypal genetic switch — the lambda phage decision circuit — have shown it to be robust against changes in binding affinity of promoters and repressors; its stable switching action arises from the structure of its network, rather than the specific affinities of its binding site²⁶. Additionally, a number of networks for biological oscillations and transcriptional regulations have been shown to be tolerant against noise (ref. 27; and see review in this issue by Rao and colleagues, pages 231–237). But only computer simulation could have shown the degree to which the gene regulatory networks for segmentation during *Drosophila* embryogenesis remain robust over a large range of kinetic parameters^{28,29}.

The robustness of a system is not always to an organism's advantage. Cancer cells are extremely robust for their own growth and survival against various perturbations. They continue to proliferate, driven by the engine of the cell cycle, eliminating

communication with their external environment, thus making it insensitive against external perturbations. In addition, many anticancer drugs are rendered ineffective by the normal functioning of a patient's body, including defence systems such as the metabolism of xenobiotics (most notably by cytochrome P450), the brain–blood barrier, and the dynamics of gene regulatory circuits, which can adjust the concentration of drug targets through feedback mechanisms and redundancy. To establish treatments that move patients from a stable but diseased state to a healthy one will require an in-depth, system-level understanding of biological robustness.

Although the general principles of robust systems are well established, there remain a number of unresolved issues concerning their evolution and execution in specific biological systems, and how they can be manipulated or designed. Control theory has been used to provide a theoretical underpinning of some robust systems, such as adaptation through negative feedback²¹. However, this approach has limitations. For example, current control theory assumes that target values or statuses are provided initially for the systems designer, whereas in biology such targets are created and revised continuously by the system itself. Such self-determined evolution is beyond the scope of current control theory.

No free lunch

Although robustness is critical in assuring the survival of a biological system, it does not come without cost. Carlson and Doyle emphasize the “robust, yet fragile” nature of complex systems exhibiting highly optimized tolerance^{30,31}. Systems designed or evolved to be robust against common or known perturbations can often be fragile to new perturbations.

Another view on the vulnerability of complex network comes from a statistical perspective^{32–34}. Comparative studies on robustness of large-scale networks show that scale-free networks (also known as ‘small world’ or Erdős–Rényi networks) are more robust than randomly connected networks against random failure of their components³⁴. However, scale-free networks are more vulnerable against malfunction of the few highly connected nodes that function as hubs.

Scale-free networks can form by growth such that new nodes are connected preferentially to nodes that are already highly connected. Barabasi and colleagues claim that protein–protein interaction networks, which constitute the protein universe (see review in this issue by Koonin and colleagues, pages 218–223), are scale-free^{32,35} and that mutations in highly connected proteins are more likely to be lethal than are mutations in less-connected nodes³³. Although they estimated connectivity from yeast two-hybrid data, which are notoriously noisy, this hypothesis is intuitively attractive. For example, the p53 protein is one of the most connected hubs in the protein universe, and its mutations cause serious damage to cellular functions, particularly in repair of DNA damage and tumour suppression³⁶.

Nevertheless, some of the claims for scale-free networks are still controversial³⁷, and evidence for mechanisms leading to preferential attachment in biological systems remains equivocal. Furthermore, yeast two-hybrid assays produce many false-positive outcomes, and the current hand-crafted pathway maps may be heavily biased towards connection to functionally important genes simply because these have been popular targets for research.

Even when these shortcomings are surpassed, such statistics-based theories — despite providing insights on macroscopic properties of the network — will still have difficulty making predictions about specific interactions. It is analogous to telling a stock-market investor that “one in 50 companies will go bankrupt”, advice that is of little help if you are unable to identify which one. The challenge for statistical theories is to identify how they can be linked to specific behaviours and so make useful predictions.

Design patterns of functional modules

Just as the principles behind robust networks can be classified into several types, so too can the various functional circuits or modules

from which they are assembled, such as genetic switches, flip-flops, logic gates, amplifiers and oscillators. Good examples come from the mechanisms of biochemical oscillations (see review in this issue by Goldbeter, pages 238–245), which have been the focus of numerous groups^{38–41}. These studies have facilitated their classification into several schemes, such as substrate-depletion oscillators, positive feedback loops, the Goodwin oscillator and time-delayed negative feedback oscillators⁴¹. Similar attempts have also been made for other functional networks. Jordan and colleagues have identified various examples of multitasking in signal transduction⁴²; Bhalla and Iyengar reported several circuits that may function as temporal information stores (that is, memory devices)⁴³; and Rao and colleagues have uncovered several circuits that mitigate the effect of noise and exploit it for specific functions (see review in this issue, pages 231–237).

Although these functional networks have analogues in electronic and process engineering, they have been formed by evolution, which makes it unlikely that any kind of ‘first principle’ underlies their design. However, a set of principles can be envisaged and identified through studying the structure and function of biological circuits, and their origin at the system level^{44–46}. What are their basic functional building blocks? What are their dynamical properties and operating principles? How has each module evolved? And how can they be adapted or designed for alternative applications?

Recently, a systematic, high-throughput computational study was carried out by Shen-Orr and colleagues, which identified common motifs in the gene regulatory networks of *E. coli* using the RegulonDB database⁴⁷. They found that feed-forward loops, single-input modules and dense overlapping regulons appeared frequently. While this study only used a gene regulation database, this type of approach can be augmented to include protein–protein and protein–DNA interactions to systematically identify network design patterns from large-scale data.

Such data, combined with function-driven identification of circuit patterns, will allow the creation of a large repository of functional biological networks, so enabling the systematic analysis of design patterns and their evolution. We already know of cases where the same circuit patterns and homologous genes produce similar system behaviours, but with unrelated physiological outcomes. We also know of cases where the same circuit patterns use different sets of genes to attain similar system behaviours, and where identical functions are achieved with degenerate paths involving different circuit patterns and different genes⁴⁶. More systematic surveys will be needed to determine how many evolutionary conserved circuits exist, in what functions and how they relate to the evolution of genes. It may be that functional circuits should be considered the units of evolution.

Systems drug and treatment discovery

The systems biology approach, with its combination of computational, experimental and observational enquiry, is highly relevant to drug discovery and the optimization of medical treatment regimes for individual patients. Although the analysis of individual single nucleotide polymorphisms is expected to reveal individual genetic susceptibilities to all forms of pathological condition, it may be impossible to identify such relationships when complex interactions are involved.

Consider a hypothetical example where variations of gene A induce a certain disease. Susceptibility relationships may not be apparent if circuits exist to compensate for the effects of the variability. Polymorphisms in gene A will be linked to disease susceptibility only if these compensatory circuits break down for some reason. A more mechanistic, systems-based analysis will be necessary to elucidate more complex relationships involving multiple genes that may create new opportunities for drug discovery and treatment optimization.

Computer simulation and analysis, along with traditional bioinformatics approaches, have frequently been proposed to significantly increase the efficiency of drug discovery^{48–50}. At present, empirical ADME/Tox (absorption distribution metabolism excretion/toxicity) and pharmacokinetic predictions have been used with some success.

For example, a human intestinal absorption model based on correlations between the passive permeation measurement of over 300 compounds and known structural features, such as hydrogen-bond donors, hydrogen-bond acceptors and molecular weight, has been used to predict the absorption of novel compounds by the human intestine⁵¹. However, such models are not easily converted for use in other situations and they often require extensive data sets in order to address specific questions. What is needed are reliable, mechanism-based ADME/Tox and pharmacokinetic models^{52–56}, built on molecular-level models of cells, that are more easily transferable and accountable than are traditional, empirical, quantitative structure–activity relations.

Scaling up

So far, most systems biology simulations have tended to target relatively small sub-networks within cells, such as the feedback circuit for bacteria chemotaxis^{20,21}, the circadian rhythm^{57,58}, parts of signal-transduction pathways^{43,59}, simplified models of the cell cycle^{7,60,61} and red blood cells^{62–64}. Notable larger simulations have attempted to model bacterial metabolic networks for analysis of metabolic control^{62,63} and flux balance^{8,65}, but these deal with steady-state rather than dynamic behaviour. Recently, research has begun on larger-scale simulations. At the level of the biochemical network, simulation of the epidermal growth factor (EGF) signal-transduction cascade has been carried out. The simulation involves over 100 equations and kinetic parameters and will be used to predict complex behaviours of the pathway, as well as to identify roles of external and internal EGF receptors⁵⁹. The physiome project is an ambitious attempt to create virtual organs that represent essential features of organs *in silico*^{66,67}. Simulation of the heart was one of the early attempts in this direction, integrating multiple scales of models from genetics to physiology⁶⁸. Even whole-patient models for specific disease, such as obesity and diabetes, are being developed for prediction of disease development and drug discovery.

Building a full-scale patient model, or even a whole-cell or organ model, is a challenging enterprise. Multiple aspects of biological processes have to be integrated and the model predictions must be verified by biological and clinical data, which are at best sparse for this purpose. Integrating heterogeneous simulation models is a non-trivial research topic by itself, requiring integration of data of multiple scales, resolutions and modalities.

Simulation often requires integration of multiple hierarchies of models that are orders of magnitude different in terms of scale and qualitative properties (for example, gene regulations, biochemical networks, intercellular communications, tissue, organ and patient). Although some processes can be modelled by either stochastic computation or differential equations alone, many require a combination of both methods. But some biochemical processes take place within a millisecond whereas others can take hours or days. Additionally, biological processes often involve the interaction of different types of process, such as biochemical networks coupled to protein transport, chromosome dynamics, cell migration or morphological changes in tissues. Although biochemical networks may be reasonably modelled using differential equations and stochastic simulation, many cell biological phenomena require calculation of structural dynamics, deformation of elastic bodies, spring-mass models and other physical processes.

Nevertheless, development of precision models and their applications to ADME/Tox models are expected to revolutionize the process of drug discovery by providing a capability for multiple-target identification and high-throughput virtual screening of compounds. Furthermore, target identification using cellular models may provide desirable structures for candidate compounds by applying multiple constraints to parallel virtual screening⁵⁴, rationalizing drug discovery into a more systematic process (Fig. 1).

Systems therapy

Surpassing its scope for efficient improvements in the current paradigm of drug discovery and treatment, the introduction of a

system-oriented view may drastically change the way treatments are conducted. Two somewhat speculative scenarios illustrate these opportunities.

Consider a feedback compensation circuit involving a drug target protein. Changes in the concentration of the protein resulting from drug administration may be neutralized by feedback control. High dosages of drugs will need to be administered to overcome this compensation mechanism, but this could produce serious side effects. Alternatively, small dosages of drugs could mitigate the feedback mechanism, so that the effect on the target protein will not be neutralized. Considering the p53 system, if there is abnormal overexpression of MDM2 (a protein that regulates p53), simply increasing p53 transcription may not restore the system to normal, as the excessive MDM2 protein will quickly ubiquitinate p53, targeting it for destruction. Additionally, p53 itself transactivates MDM2. MDM2 activity must be suspended or reduced to a normal level, at least temporarily, to make p53 stimulation effective in inducing cell-cycle arrest or apoptosis. The highly effective administration of multiple drug regimes can be accomplished only with a system-level analysis of the dynamics of gene regulatory circuits.

A far more futuristic approach proposes the introduction of functional genetic circuits to control cellular dynamics *in vivo* (see review in this issue by Hasty and colleagues, pages 224–230). Already, a set of basic functional circuits, such as oscillators and toggle switches, has been constructed and its viability confirmed in *E. coli* (refs 69–71; and see review by Hasty and colleagues). Computer simulation and comprehensive analysis will be needed to ensure that such circuits function as intended and do not result in significant side-effects. In the future, perhaps a genetic circuit can be devised to sense the level of p53 protein when DNA is damaged and switch on circuits to further increase transcription of p53.

The application of systems biology to medical practice is the future of medicine. Its realization will see drug discovery and the design of multiple drug therapies and therapeutic gene circuits being pursued just as occurs now with modern, complex engineering products — through iterative cycles of hypothesis and simulation-driven processes (Fig. 1). Although the road ahead is long and winding, it leads to a future where biology and medicine are transformed into precision engineering. □

doi:10.1038/nature01254

- Baldi, P. & Brunak, S. *Bioinformatics: The Machine Learning Approach* 2nd edn (MIT Press, Cambridge, MA, 2001).
- Onami, S., Kyoda, K., Morohashi, M. & Kitano, H. in *Foundations of Systems Biology* (ed. Kitano, H.) 59–75 (MIT Press, Cambridge, MA, 2001).
- Ideker, T. E., Thorsson, V. & Karp, R. M. in *Pac. Symp. Biocomput.* (eds Altman, R. B., Dunker, A. K., Hunter, L., Lauderdale, K. & Klein, T. E.) 305–316 (World Scientific, Singapore, 2000).
- Ideker, T. *et al.* Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* **18**(Suppl. 1), S233–S240 (2002).
- Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934 (2001).
- Borisuk, M. T. & Tyson, J. J. Bifurcation analysis of a model of mitotic control in frog eggs. *J. Theor. Biol.* **195**, 69–85 (1998).
- Chen, K. C. *et al.* Kinetic analysis of a molecular model of the budding yeast cell cycle. *Mol. Biol. Cell* **11**, 369–391 (2000).
- Edwards, J. S., Ibarra, R. U. & Palsson, B. O. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nature Biotechnol.* **19**, 125–130 (2001).
- Fell, D. *Understanding the Control of Metabolism* (Portland, London, 1997).
- Morohashi, M. *et al.* Robustness as a measure of plausibility in models of biochemical networks. *J. Theor. Biol.* **216**, 19–30 (2002).
- Kitano, H. Standards for modeling. *Nature Biotechnol.* **20**, 337 (2002).
- Hucka, M. *et al.* in *Pac. Symp. Biocomput.* (eds Altman, R. B., Dunker, A. K., Hunter, L. & Klein, T. E.) 450–461 (World Scientific, Singapore, 2002).
- Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
- Alliance for Cellular Signaling <<http://www.Afcs.org/>> (2002).
- Signal Transduction Knowledge Environment <<http://www.stke.org/>> (2002).
- Wiener, N. *Cybernetics: Or Control and Communication in the Animal and the Machine* (MIT Press, Cambridge, MA, 1948).
- Bertalanffy, L. v. *General System Theory* (Braziller, New York, 1968).
- Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662–1664 (2002).
- Kitano, H. in *Foundations of Systems Biology* (ed. Kitano, H.) 1–36 (MIT Press, Cambridge, MA, 2001).
- Alon, U. *et al.* Robustness in bacterial chemotaxis. *Nature* **397**, 168–171 (1999).

21. Yi, T. M. *et al.* Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proc. Natl Acad. Sci. USA* **97**, 4649–4653 (2000).
22. Barkai, N. & Leibler, S. Robustness in simple biochemical networks. *Nature* **387**, 913–917 (1997).
23. Weng, G., Bhalla, U. S. & Iyengar, R. Complexity in biological signaling systems. *Science* **284**, 92–96 (1999).
24. Levine, K., Tinkelenberg, A. & Cross, F. in *Progress in Cell Cycle Research* (eds Meijer, L., Guidet, S. & Lim Tung, H. Y.) 101–114 (Plenum, New York, 1995).
25. Chang, D. E., Smalley, D. J. & Conway, T. Gene expression profiling of *Escherichia coli* growth transitions: an expanded stringent response model. *Mol. Microbiol.* **45**, 289–306 (2002).
26. Little, J. W., Shepley, D. P. & Wört, D. W. Robustness of a gene regulatory circuit. *EMBO J.* **18**, 4299–4307 (1999).
27. Gonze, D., Halloy, J. & Goldbeter, A. Robustness of circadian rhythms with respect to molecular noise. *Proc. Natl Acad. Sci. USA* **99**, 673–678 (2002).
28. von Dassow, G. *et al.* The segment polarity network is a robust developmental module. *Nature* **406**, 188–192 (2000).
29. Eldar, A. *et al.* Robustness of the BMP morphogen gradient in *Drosophila* embryonic patterning. *Nature* **419**, 304–308 (2002).
30. Carlson, J. M. & Doyle, J. Highly optimized tolerance: a mechanism for power laws in designed systems. *Phys. Rev. E* **60**, 1412–1427 (1999).
31. Carlson, J. M. & Doyle, J. Complexity and robustness. *Proc. Natl Acad. Sci. USA* **99**, 2538–2545 (2002).
32. Jeong, H. *et al.* The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
33. Jeong, H. *et al.* Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
34. Albert, R., Jeong, H. & Barabási, A. L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
35. Podani, J. *et al.* Comparable system-level organization of Archaea and Eukaryotes. *Nature Genet.* **29**, 54–56 (2001).
36. Vogelstein, B., Lane, D. & Levine, A. J. Surfing the p53 network. *Nature* **408**, 307–310 (2000).
37. Adamic, L. A., Lukose, R. M., Puniyani, A. R. & Huberman, B. A. Search in power-law networks. *Phys. Rev. E* **64**, 046135–1–046135–8 (2001).
38. Higgins, J. The theory of oscillating reactions. *Ind. Eng. Chem.* **59**, 18–62 (1967).
39. Berridge, M. J. & Rapp, P. E. A comparative survey of the function, mechanism and control of cellular oscillators. *J. Exp. Biol.* **81**, 217–279 (1979).
40. Goldbeter, A. *Biochemical Oscillations and Cellular Rhythms* (Cambridge Univ. Press, Cambridge, 1996).
41. Tyson, J. J. in *Computational Cell Biology* (eds Fall, C. P., Marland, E. S., Wagner, J. M. & Tyson, J. J.) 230–260 (Springer, New York, 2002).
42. Jordan, J. D., Landau, E. M. & Iyengar, R. Signaling networks: the origins of cellular multitasking. *Cell* **103**, 193–200 (2000).
43. Bhalla, U. S. & Iyengar, R. Emergent properties of networks of biological signaling pathways. *Science* **283**, 381–387 (1999).
44. Hartwell, L. H. *et al.* From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
45. Csete, M. E. & Doyle, J. C. Reverse engineering of biological complexity. *Science* **295**, 1664–1669 (2002).
46. Edelman, G. M. & Gally, J. A. Degeneracy and complexity in biological systems. *Proc. Natl Acad. Sci. USA* **98**, 13763–13768 (2001).
47. Shen-Orr, S. S. *et al.* Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genet.* **31**, 64–68 (2002).
48. Cascante, M. *et al.* Metabolic control analysis in drug discovery and disease. *Nature Biotechnol.* **20**, 243–249 (2002).
49. Bailey, J. E. Lessons from metabolic engineering for functional genomics and drug discovery. *Nature Biotechnol.* **17**, 616–618 (1999).
50. Bailey, J. E. Reflections on the scope and the future of metabolic engineering and its connections to functional genomics and drug discovery. *Metab. Eng.* **3**, 111–114 (2001).
51. Lipinski, C. A. *et al.* Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **46**, 3–26 (2001).
52. Butina, D., Segall, M. D. & Frankcombe, K. Predicting ADME properties in silico: methods and models. *Drug Discov. Today* **7**, S83–S88 (2002).
53. Ekins, S. & Rose, J. In silico ADME/Tox: the state of the art. *J. Mol. Graph. Model.* **20**, 305–309 (2002).
54. Selick, H. E., Beresford, A. P. & Tarbit, M. H. The emerging importance of predictive ADME simulation in drug discovery. *Drug Discov. Today* **7**, 109–116 (2002).
55. Li, A. P. & Segall, M. Early ADME/Tox studies and in silico screening. *Drug Discov. Today* **7**, 25–27 (2002).
56. Ekins, S. *et al.* Progress in predicting human ADME parameters in silico. *J. Pharmacol. Toxicol. Methods* **44**, 251–272 (2000).
57. Ueda, H. R., Hagiwara, M. & Kitano, H. Robust oscillations within the interlocked feedback model of *Drosophila* circadian rhythm. *J. Theor. Biol.* **210**, 401–406 (2001).
58. Leloup, J. C., Gonze, D. & Goldbeter, A. Limit cycle models for circadian rhythms based on transcriptional regulation in *Drosophila* and *Neurospora*. *J. Biol. Rhythms* **14**, 433–448 (1999).
59. Schoeberl, B. *et al.* Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nature Biotechnol.* **20**, 370–375 (2002).
60. Tyson, J. J. & Novak, B. Regulation of the eukaryotic cell cycle: molecular antagonism, hysteresis, and irreversible transitions. *J. Theor. Biol.* **210**, 249–263 (2001).
61. Novak, B. *et al.* Mathematical model of the fission yeast cell cycle with checkpoint controls at the G1/S, G2/M and metaphase/anaphase transitions. *Biophys. Chem.* **72**, 185–200 (1998).
62. Ni, T. C. & Savageau, M. A. Model assessment and refinement using strategies from biochemical systems theory: application to metabolism in human red blood cells. *J. Theor. Biol.* **179**, 329–368 (1996).
63. Ni, T. C. & Savageau, M. A. Application of biochemical systems theory to metabolism in human red blood cells. Signal propagation and accuracy of representation. *J. Biol. Chem.* **271**, 7927–7941 (1996).
64. Jamshidi, N. *et al.* Dynamic simulation of the human red blood cell metabolic network. *Bioinformatics* **17**, 286–287 (2001).
65. Edwards, J. S. & Palsson, B. O. Robustness analysis of the *Escherichia coli* metabolic network. *Biotechnol. Prog.* **16**, 927–939 (2000).
66. Bassingthwaite, J. B. Strategies for the physiome project. *Ann. Biomed. Eng.* **28**, 1043–1058 (2000).
67. Rudy, Y. From genome to physiome: integrative models of cardiac excitation. *Ann. Biomed. Eng.* **28**, 945–950 (2000).
68. Noble, D. Modeling the heart—from genes to cells to the whole organ. *Science* **295**, 1678–1682 (2002).
69. Guet, C. C. *et al.* Combinatorial synthesis of genetic networks. *Science* **296**, 1466–1470 (2002).
70. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403**, 339–342 (2000).
71. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–338 (2000).

Acknowledgements

I thank S. Imai, J. Doyle, J. Tyson, T.-M. Yi, N. Hiroi and M. Morohashi for their useful comments on the manuscript. This research is, in part, supported by: the Rice Genome and Simulation Project (Ministry of Agriculture), International Standard Development area of International Joint Research Grant (New Energy and Industrial Technology Development Organization (NEDO)/Japanese Ministry of Economy, Trade and Industry (METI)), Exploratory Research for Advanced Technology (ERATO) and Institute for Bioinformatics Research and Development (BIRD) program (Japan Science and Technology Corporation), and through the special coordination funds for promoting science and technology from the Japanese government's Ministry of Education, Culture, Sports, Science, and Technology.

The language of genes

David B. Searls

Bioinformatics Division, Genetics Research, GlaxoSmithKline Pharmaceuticals, 709 Swedeland Road, PO Box 1539, King of Prussia, Pennsylvania 19406, USA (e-mail: david_b_searls@gsk.com)

Linguistic metaphors have been woven into the fabric of molecular biology since its inception. The determination of the human genome sequence has brought these metaphors to the forefront of the popular imagination, with the natural extension of the notion of DNA as language to that of the genome as the 'book of life'. But do these analogies go deeper and, if so, can the methods developed for analysing languages be applied to molecular biology? In fact, many techniques used in bioinformatics, even if developed independently, may be seen to be grounded in linguistics. Further interweaving of these fields will be instrumental in extending our understanding of the language of life.

The science of linguistics has fully as many facets and fields as biology, and like biology, what may be called its 'modern era' can be traced to the 1950s¹. The decade that unveiled the structure of DNA also witnessed a revolution in linguistics led by Noam Chomsky, whose work radically diversified the field beyond its then-current focus on simply cataloguing the actual utterances of a language, to exploring the mechanisms by which they are produced. Seeking to identify the universals at the core of all languages, he posited a new, generative form of grammar, or set of syntactic rules, that would help to account for the immense creativity in the production of language that emerges so rapidly as individuals develop².

In pursuit of his 'universal grammar', Chomsky created waves that washed up on many scientific shores. Besides his profound influence on theoretical linguistics, his mathematical approach to the description of languages prompted a burst of development in formal language theory. This produced methods with widespread utility in computer science, from the specification and interpretation of computer languages to the fields of syntactic pattern recognition, natural language processing and speech understanding³. The Chomsky hierarchy of language classes has proven especially durable as a means of stratifying formal languages according to their expressive power and resulting computational and mathematical complexity (Box 1). Chomsky's influence has also extended to cognitive science, analytic philosophy and even literary criticism. The common experience in a number of fields is that it is not only analytic techniques derived from linguistics, but also what might be called a linguistic sensibility, that can illuminate and inform other similarly complex domains.

Mathematical linguistics and macromolecules

In the 1980s, several workers began to follow various threads of Chomsky's legacy in applying linguistic methods to molecular biology. Early results included the fundamental observation that formal representations could be applied to biological sequences⁴ — the extension of linguistic formalisms in new, biologically inspired directions⁵ — and the demonstration of the utility of grammars in capturing not only informational but also structural aspects of macromolecules⁶.

Nucleic acid linguistics

From this work there followed a series of mathematical results concerning the linguistics of nucleic acid structure^{7–9}. These results derive from the fact that a folded RNA secondary

structure entails pairing between nucleotide bases that are at a distance from each other in the primary sequence, establishing relationships that in linguistics are called dependencies. The most basic secondary-structure element is the stem-loop, in which the stem creates a succession of nested dependencies that can be captured in idealized form by the following context-free base-pairing grammar⁷ (Box 1):

$$S \rightarrow gSc \quad S \rightarrow cSg \quad S \rightarrow aSu \quad S \rightarrow uSa \quad S \rightarrow \varepsilon$$

(The ε in the last rule indicates that an S is simply erased.) This grammar affords any and every derivation of 'hairpin' sequences of a form such as the following:

$$S \rightarrow gSc \Rightarrow gaSuc \Rightarrow gauSauc \Rightarrow \dots \Rightarrow gaucgaSucgauc \Rightarrow gaucgaucgauc$$

Derivations from this grammar grow outward from the central S , creating the nested dependencies of the stem (Fig. 1a), analogous to such phenomena as nested relative clauses in natural language (for example, "The gene that the scientist whom our grant supported discovered encoded a kinase"). In a realistic stem-loop, the derivation would terminate in an unpaired loop of at least several bases and might also contain, for example, non-Watson-Crick base pairs and 'bulges'. But such features are easily added to the grammar without affecting the fundamental result that any language consisting of RNA sequences that fold into these basic structures requires context-free expression¹⁰.

In addition to stem-loop structures, arbitrarily branched folded structures may be captured by simply adding to the grammar above a rule $S \rightarrow SS$, whose application creates bifurcations in the derivation tree⁷ (Fig. 1b). The base-pairing dependencies remain non-crossing, although more complicated. The resulting grammar is formally ambiguous, meaning that there are guaranteed to be sequences in the language for which more than one derivation tree is possible¹⁰. Thus, the string *gaucgaucgauc* can be derived as a single hairpin or as a branched structure (Fig. 1a, b). This linguistic property of ambiguity, reflected in natural languages in sentences that can be syntactically parsed in more than one way (for example, "She saw the man with the telescope"), directly models the biological phenomenon of alternative secondary structure⁷. Although these models are only abstractions of a thermodynamically determined process, ambiguity allows them to embody the ensemble of potential secondary structures, and more specific grammars can specify particular forms, such as transfer RNA cloverleafs⁹.

Box 1

The Chomsky hierarchy and formal language theory

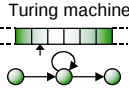
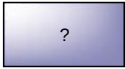
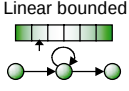
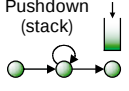
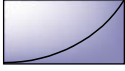
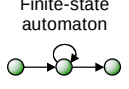
Formal language theory defines languages to be nothing more than sets of strings of symbols drawn from some alphabet. A grammar is a rule-based approach to specifying a language, consisting of a set of rewriting rules that take forms such as $A \rightarrow xB$. Here, upper-case letters denote temporary or nonterminal symbols, which do not occur in the alphabet, whereas lower-case letters are terminal symbols that do. The example rule specifies that any occurrence of the nonterminal A may be replaced by an x followed by a B .

Beginning with a starting nonterminal S , a derivation from a grammar consists of a series of rewriting steps that ends when the last nonterminal is eliminated. Consider the simple grammar with an alphabet x and y , and containing the rules $S \rightarrow xS$ and $S \rightarrow y$. This grammar generates all strings beginning with any number of x 's and ending in a single y . It produces derivations such as $S \rightarrow xS \Rightarrow xxS \Rightarrow xxxS \Rightarrow xxxxy$, where each double arrow signifies the application of a single-arrow rule. In this case there are three applications of the first rule followed by a single application of the second to produce a terminal string, one of the infinite number of such strings in this language.

Any grammar whose rules rewrite a nonterminal as a terminal followed by at most one nonterminal is called regular, and is said to generate a regular language. An equivalent means of generating such languages is a finite-state automaton (FSA), a notional machine used to reason about computation, built out of states (circles; see figure opposite) which are interconnected by transitions (arrows) that emit symbols from the alphabet as they are traversed.

Grammars that allow any arrangement of terminals and nonterminals on the right-hand sides of rules have greater expressive power. They are called context-free grammars, and can generate not only all regular languages, but also non-regular languages such as strings of x 's followed by the same number of y 's (for example, $xxxxxyyy$). Such languages cannot be specified by a regular grammar or FSA because these devices have no mechanism for 'remembering' how many x 's were generated when the time comes to derive the y 's. This shortcoming is remedied by means of context-free rules such as $S \rightarrow xSy$, which always generate an x and a y at the same time. Alternatively, an automaton augmented with a push-down store, a memory device that pushes or pops symbols to or from a stack during transitions, also provides such a counting capability. In either case, context-free languages allow strings that embody dependencies between terminals, such as the relationship matching x 's and y 's in the example, provided that those dependencies can be drawn as nested, either strictly within or independent of each other, but never crossing.

Even context-free grammars are inadequate for some languages, for instance strings of consecutive x 's, y 's and z 's in equal number (for

Language	Automaton	Grammar	Recognition
Recursively enumerable languages		Unrestricted $Baa \rightarrow A$	Undecidable 
Context-sensitive languages		Context sensitive $At \rightarrow aA$	Exponential? 
Context-free languages		Context free $S \rightarrow gSc$	Polynomial 
Regular languages		Regular $A \rightarrow cA$	Linear 

example, $xxxxxyyyzzz$). This entails dependencies that necessarily cross one another, called cross-serial dependencies, and to capture these with a grammar requires rules that have additional symbols on their left-hand side (though never more than on their right-hand side). Such context-sensitive rules correspond to automata with a more sophisticated memory device, a tape whose length is bounded in a certain way, upon which the machine can read and write symbols. Context-sensitive languages include all context-free languages and many more, yet theoretically there exist languages outside even this set, called recursively enumerable languages, generated by grammars of completely unrestricted form or by machines with unbounded tapes best known as Turing machines.

In the figure above, the language classes in the left column contain exactly those languages that can be generated by the automata and grammar types indicated in the next two columns. Each level contains all of those below it. The right-hand column illustrates the computational complexity, in the general case, of recognizing whether a string belongs in a given language, showing how the time required grows as a function of the length of the input string. At the highest level of the hierarchy, one is not even guaranteed to be able to arrive at an answer by computational means. This is just one indication of the trade-off between the increase in expressive power afforded by ascending the Chomsky hierarchy, and the mathematical and algorithmic limitations that invariably result²⁴.

Finding that the language of RNA is at least context-free has mathematical and computational consequences, for example, for the nature and inherent performance bounds of any algorithm dealing with secondary structure (Box 1). For instance, the fast, regular-expression search tools used commonly in bioinformatics (such as those in the popular Perl scripting language) are ruled out, as in their standard form they specify only regular languages. These consequences show the importance of characterizing linguistic domains in the common terminology and methodology of formal language theory, so as to connect them immediately to the wealth of tools and understanding already available. For this reason, recent bioinformatics textbooks have devoted whole chapters to the relationship of biological sequences to the Chomsky hierarchy^{11,12}.

In light of these practical consequences of linguistic complexity, a significant finding is that there exist phenomena in RNA that in fact raise the language even beyond context-free. The most obvious of

these are so-called non-orthodox secondary structures such as pseudoknots, which are pairs of stem-loop elements in which part of one stem resides within the loop of the other (Fig. 1c). This configuration induces cross-serial dependencies in the resulting base pairings, requiring context-sensitive expression (Box 1). Predictably, given this further promotion in the Chomsky hierarchy, the need to encompass pseudoknots within secondary-structure recognition and prediction programs has significantly complicated algorithm design¹³. Another non-context-free phenomenon that occurs in RNA is a consequence of alternative secondary structure, such as that seen in bacterial attenuators, which are regulatory elements that depend on switching between conformations in nascent mRNA molecules. For any grammar required to simultaneously represent both conformations, these mutually exclusive options create overlapping (and thus cross-serial) dependencies in the alternate base-pairing schemes⁷ (Fig. 1d).

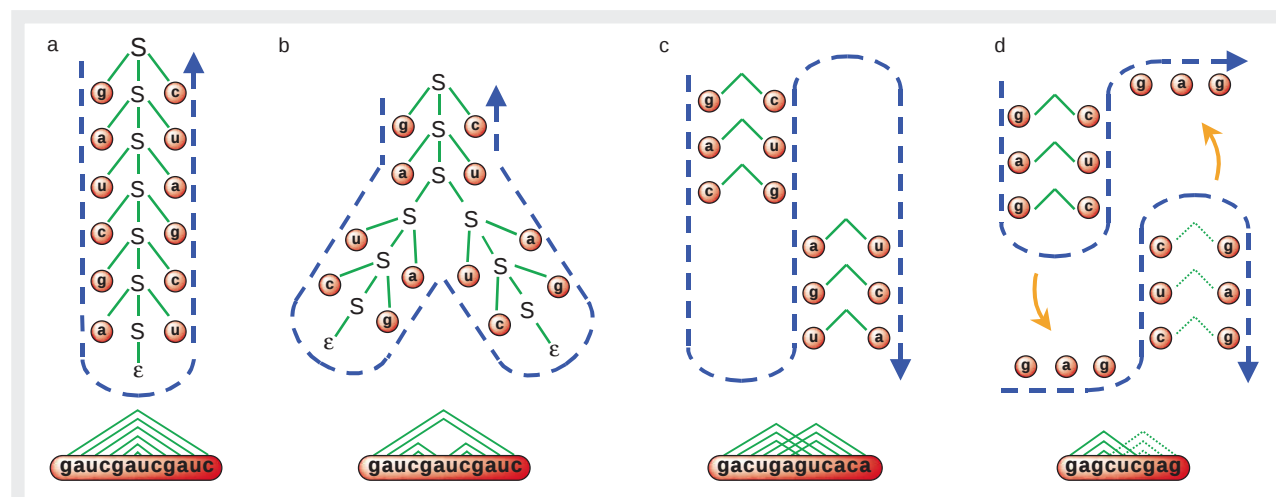


Figure 1 Grammar-style derivations of idealized versions of RNA structures. **a**, A stem; **b**, a branched structure; **c**, a pseudoknot; and **d**, alternative secondary structures of an attenuator. The trees for **a** and **b** are graphical depictions of derivations from grammars given in the text. By convention, a starting nonterminal *S* is at the root of the tree and gives rise to branches for each symbol to which it rewrites in the course of the derivation. The string derived can be read by tracing the frontier or leaf nodes of the tree, left to right (dashed blue lines). For **c** and **d**, derivation trees are not explicitly indicated because of the complexity of the context-sensitive grammars required⁷. The same strings are also shown in linear fashion, with dependencies indicated between terminals derived at the same steps.

Using formalisms called tree-adjoining grammars and their variants¹⁴, which are considered to be mildly context-sensitive and relatively tractable, it is possible to encompass a wide range of RNA secondary structures¹⁵. Additionally, new types of grammars have been invented to deal with such biological examples^{16,17}. Natural languages seem to be beyond context-free as well, based on linguistic phenomena entailing cross-serial dependencies¹⁸, although in both domains such phenomena seem to be less common than nested dependencies. Thus, by one measure at least, nucleic acids may be said to be at about the same level of linguistic complexity as natural human languages.

Protein linguistics

There has been less activity in modelling proteins with linguistic methods, perhaps because they are viewed as having a richer basic repertoire of interactions and conformations than nucleic acids, and perhaps also more of a sense of emergent properties. Yet grammars can be extraordinarily detailed and nuanced (while remaining manageable because of their inherently modular and hierarchical design), and moreover need not capture every aspect of a structure to be useful. In fact, the comprehensiveness and proper role of grammars remains as much an issue for natural language as it might prove to be for proteins, as does the question of whether exemplars of either language are susceptible of a compositional semantics (that is, one for which the meaning or function of the whole can be built up in rule-based fashion from that associated with its parts)³. In any case there is a decidedly linguistic flavour to certain abstracted depictions of protein structure, such as domain schematics (for example, the SMART system, which portrays the highly variable arrangements of 'mobile' domains¹⁹) or topology 'cartoons' (for example, the TOPS system, which annotates dependencies between secondary structural elements, including positional and chiral relationships²⁰).

Specific aspects of protein structure have been modelled explicitly with grammars. Secondary structural elements, and in particular the hydrogen bonding between strands in a β -sheet, may be arrayed in antiparallel fashion, creating nested dependencies by analogy with stem-loop structures in RNA, or in parallel fashion, which creates cross-serial dependencies. Such arrangements have been represented using stochastic tree grammars²¹, which are related to tree-adjoining grammars and which have also been shown to generate a range of configurations of β -sheets that corresponds well to that seen in nature (A. Joshi, personal communication). Another grammar-

based approach, using tools from graph theory, was shown recently to be capable of generating a preponderance of the class of all- β -folds from just four basic rules²².

Mathematicians are concerned with closure properties of languages, that is, whether they remain at the same level of the Chomsky hierarchy when various operations are performed on their contents⁹. Simple concatenation of strings is a so-called regular operation, whereas insertion of one string in another is a context-free operation, insofar as it never causes dependencies to cross, but only further nests them. Neither operation raises a context-free language beyond context-free, nor (it can be shown) do a series of biological operations such as replication and recombination¹⁰. However, translocation of segments of a string may create cross-serial dependencies where none existed before, and thus the block movements typical of genomic rearrangements may constitute an upward force in the Chomsky hierarchy that is inherent in evolution¹⁰.

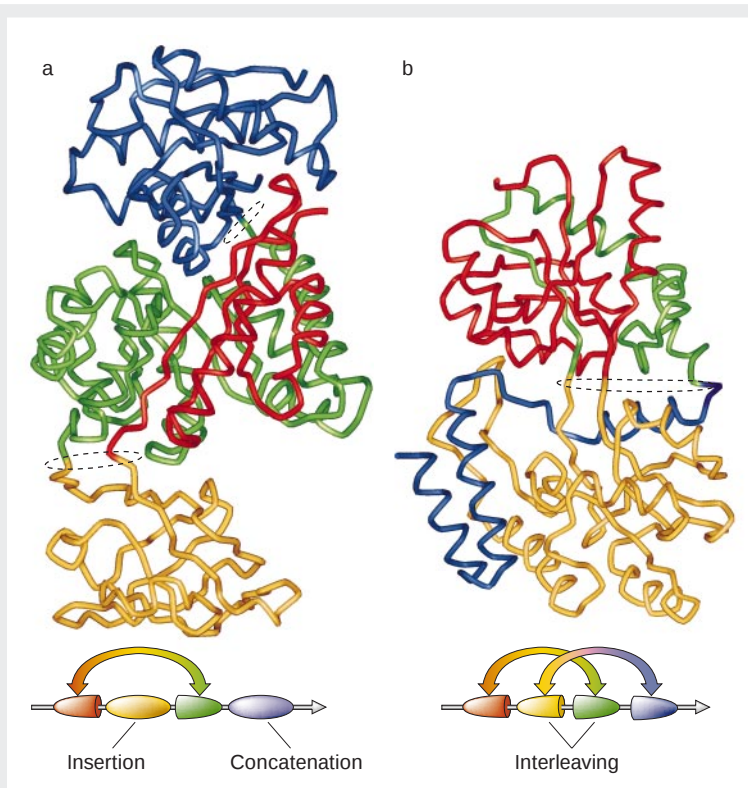
Nevertheless, within proteins we see evidence that at the level of domains (if not supersecondary structure) there is again a relative scarcity of non-context-free forms (Fig. 2). This is perhaps attributable not only to the greater complexity of the genomic changes required, but also to the energetic barriers that might be anticipated in folding knot-like cross-serial dependencies, by analogy with difficulties they pose in linguistic analysis. In light of this, it is interesting that the special case of circular permutations (that is, head-to-tail rearrangements), to which protein domains seem more prone²³, do in fact preserve context-free status from a mathematical perspective²⁴.

Computational linguistics and genes

The results summarized above all relate to structural aspects of macromolecules, that is, factors inherent in their biophysical behaviour and independent of any information they contain. Yet genes do convey information, and furthermore this information is organized in a hierarchical structure whose features are ordered, constrained and related in a manner analogous to the syntactic structure of sentences in a natural language. It is thus not surprising that a number of themes, both explicit and implicit, have found their way from computational linguistics to computational biology.

One implicit theme is a convergence between organizational schemes in the two fields. Language processing is often conceived as proceeding from (1) the lexical level, at which individual words from

Figure 2 Protein domain arrangements and the Chomsky hierarchy. Shown are backbone structures for **a**, cat muscle pyruvate kinase (1pkm in Protein Data Bank; minus a short amino-terminal domain) and **b**, *Escherichia coli* D-maltodextrin binding protein (1omp in Protein Data Bank). At the bottom are schemas of the domain relationships, with double arrows connecting segments participating in the same domain. The upper, carboxy-terminal (blue) domain of 1pkm attaches by way of a simple concatenation, which is a regular operation commonly seen in proteins. The central red-and-green α/β -barrel, however, is interrupted in the middle by an insertion of the lower (orange) domain, a context-free operation insofar as it thus creates a strictly nested dependency between the divided domain segments (as would any number of domain insertions at any point). Insertions are less common than concatenations, but still fairly frequent. The two main domains of 1omp, on the other hand, seem to be interleaved, thus creating cross-serial dependencies that are necessarily context-sensitive. Whether the C-terminal (blue) segment is involved fully in the lower domain's core, however, is open to question; in any case, true interleaved structural domains seem to be very rare. The dashed ellipses in the backbone diagrams illustrate that the number of crossovers between domains (1, 2 and 3, respectively) is indicative of the level in the Chomsky hierarchy of the resulting domain arrangement.



a linear input stream (of, for example, phonemes or characters) are recognized and characterized; to (2) the syntactic level, at which words are grouped and related hierarchically according to grammar rules to form a structural description; to (3) the semantic level, at which some representation of meaning is assigned to the resulting structure, derived from that of its individual lexical elements; and finally to (4) the pragmatic level, at which language is viewed in a larger context encompassing the roles and interrelationships of sentences (and certain references within them such as pronouns) in an overall discourse or dialogue³. This progression maps neatly and meaningfully onto one used widely in biology, of sequence to structure to function to role²⁵.

In particular, the distinction between syntax and semantics (famously exemplified by Chomsky with his grammatical yet meaningless "Colourless green ideas sleep furiously"²²) is pertinent to biology. Consider two types of sequence: a string of words, and a segment of a genome. A parsing step may be seen as determining whether the words form a grammatical sentence, or, notionally, whether the genomic sequence will support the production of a polypeptide according to rules implicit in the transcriptional and translational machinery of the cell; in both cases the processes are mechanical, in fact largely processive. Then, an interpretative step determines whether the resulting sentence is meaningful, according to laws of logic and experience, or whether the polypeptide will fold into a compact core and orient its side chains so as to do useful work, a process governed by laws of thermodynamics and biochemistry. Mutated genes that are expressed but do not allow for a functional fold may be said to pass the first test but not the second.

The natural history of gene-finding algorithms offers another illustration. In the 1980s, detecting genes (in what genomic sequence was then extant) was strictly a lexical affair. Algorithms simply scanned an input sequence and within a moving window assessed its 'coding potential' on the basis of statistical measures such as oligonucleotide frequencies and periodicities. It was also possible to detect signals such as putative splice sites, again as individual lexical elements. Then, in the early 1990s, programs began to appear that

assembled lexical elements hierarchically and imposed constraints of a distinctly syntactic cast. (Thus, just as sentence constituents must agree as to number, gender, tense, and so on, so had putative exons to maintain a reading frame across whole genes.) Indeed, one program that performed creditably at the time was based explicitly on a gene grammar and a general-purpose parser (a program that determines if an input is a valid instance of any given grammar and, if so, produces a tree-structured description of the parse)²⁶.

One advantage of linguistic gene recognition was the natural accommodation of ambiguity in the form of multiple transcripts attributable, for example, to alternative splicing. Another advantage was versatility: the same parser, but with different grammars substituted, was effective in recognizing such features as tRNA genes and group I introns, including secondary structure extending to pseudoknots²⁷. Yet another area in which grammars have proven apt is in the specification of gene regulatory elements, with their highly variable distribution of disparate features. This use, in fact, was one of the first suggested biological applications of Chomsky-style grammars²⁸ and remains an active area of research^{29,30}.

Although having the advantage of flexibility, general-purpose parsers cannot compete in efficiency with programming that is customized to a particular domain, especially one that does not greatly benefit from the capacity of grammars to specify variations on a theme with ease. (English grammar would be superfluous if every sentence were patterned on the same basic declarative template.) Consequently, latter-day gene-finding algorithms, which have the 'standard model' gene structure hard-wired, do not make use of grammars *per se*. However, what has instead become a dominant technique in the analysis of biological sequences, the hidden Markov model (HMM), also traces its pedigree to linguistic roots and inherits a different set of advantages.

An HMM is a variety of automaton annotated with probability values that govern its behaviour³. They were first widely deployed in the field of speech recognition and more recently have found their way into a number of applications for the analysis of biological sequences, beginning with protein family profiles³¹. HMM

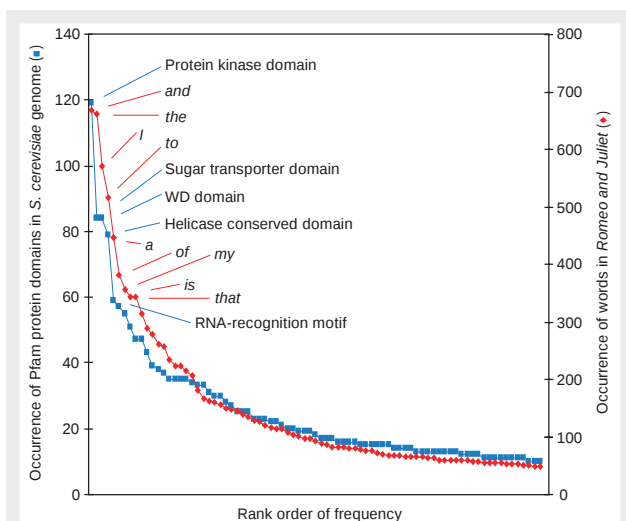


Figure 3 Distributions of the number of occurrences of Pfam protein domains (blue squares) in the genome of the yeast *Saccharomyces cerevisiae*, and of words (red diamonds) in Shakespeare's *Romeo and Juliet*, in both cases sorted in rank order from left to right. The most frequently occurring domains and words are labelled. In both cases (and in many other genomes and texts) the curves are good fits to a power-law distribution known as Zipf's law, which relates the frequency to the inverse of the rank.

architectures embody what amounts to a syntax and use an associated set of algorithms to refine and employ the model. HMMs with sophisticated domain models form the basis for several leading gene finders, including GenScan³¹ and Genie³², and the gene-finding application has driven further refinement of the method as well. The recent marked trend in computational biology towards probabilistic methods such as HMMs has mirrored a similar turn in natural language processing, which has been invigorated by a shift towards finite-state and stochastic approaches³. The use of HMMs in the two fields has been compared directly in a recent review³³.

The automata associated with HMMs are at the lowest rung of the Chomsky hierarchy and are thus inadequate for such non-regular features as the secondary structure in tRNA. This shortcoming has been addressed by adding probabilities to context-free grammars to create stochastic context-free grammars and then adapting the HMM algorithms to work with the resulting data structures³⁴. Such systems have proven useful not only in tRNA detection³⁵, but also in a variety of related biological applications^{36–39}, and have even been extended to non-context-free structures⁴⁰.

Historical linguistics and evolution

Long before Chomsky's revolution, historical linguistics was the dominant discipline in the field⁴¹, driven largely by an increasingly systematic attempt to account for the descent of modern languages from a hypothesized proto-Indo-European language first proposed in 1786¹. Of this work Darwin himself noted that "the formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel"⁴². These parallels have since inspired many authors. Dawkins' concept of 'memes' as replicating cultural fragments undergoing darwinian selection encompasses language change⁴³, as does a recent synthesis of formal language theory, learning theory and evolutionary dynamics⁴⁴. Strong analogies between the evolution of languages and of species have even formed the basis for serious scientific arguments against creationism⁴⁵. Cavalli-Sforza has comprehensively explored how population genetics can aid understanding of language evolution from a demographic perspective⁴⁶, and biological phylogenetic-reconstruction techniques have also been applied to languages⁴⁷.

Among the methods linguists themselves have used to draw 'family trees' of languages has been the statistical comparison of vocabularies, or lexicostatistics⁴¹. This approach posits that, across many languages, there is a basic, core set of cognates (essentially, word 'orthologues') relating to universal human experience and relatively resistant to change. In the 1950s, Swadesh established 200 such concepts (for example, *I*, *this*, *not*, *person*, *fish*, *blood*, *egg*, *knee*, *cloud*, *mountain* and *good*) and, based on similarity of corresponding words in different languages, derived quantitative measures of overall language relatedness⁴⁸. He further proposed that language divergences could be dated in this manner by assuming a constant rate of lexical change, a technique called glottochronology. Although controversial, this is clearly echoed in the notion of the evolutionary 'molecular clock'. Indeed, the need to account for varying rates of change in different words and proteins has been recognized independently in each field⁴⁹.

The compilation of core vocabularies from multiple languages resembles efforts to assemble 'minimal gene sets' presumed sufficient to support life (by one estimate, numbering about 300) by taking intersections of multiple genomes, and similar cautions have been noted in their use and interpretation⁵⁰. For instance, from the fact that French has no word for 'shallow' one could not conclude that the language is impoverished, any more than the apparent absence of a given enzyme necessarily rules out a certain metabolic capacity. Comparisons of gene contents across phylogeny have been used in ways that might have been drawn directly from the lexicostatistical literature. Examples include the collection of clusters of orthologous groups⁵¹ and the use of degree of overlap of gene complements (as opposed to individual sequence similarities) as a basis for phylogeny construction^{52–54} as well as a predictor of protein function⁵⁵. Both fields contend with complications introduced by synonyms and false cognates ('*faux amis*') on the one hand, and on the other, non-orthologous gene displacement and functional shifts, while recent theory concerning reticulate evolution harkens back to well-studied phenomena of language mixture such as creolization⁵⁶.

Words themselves arise and evolve by mechanisms that have been compared to biological drivers of diversity, such as mutation and recombination (called blending by linguists)⁵⁷. One mechanism they clearly have in common is compounding. The atomic units of linguistic meaning are morphemes, typically stems and affixes that combine to form words, whereas lexical units are lexemes, which may be single words or compounds and certain unitary phrases³. In like manner, proteins are considered to comprise one or more functional domains, and a recent study hypothesizes ancient 'antecedent domain segments', relating these explicitly to linguistic variation⁵⁸.

There is more than a surface similarity to such conventions, insofar as these are all elements that are surmised to combine and re-assort in the course of evolution, affording combinatorial diversity, and some of the same techniques have been applied in their analysis. For instance, a quantitative approach to the association of words is collocation analysis. Here, the frequency of co-occurrence of words in text is not only a useful heuristic in stochastic parsing³, but also provides clues in lexical semantic studies, for which compounds have been classified into such categories as noun + noun constituents, idioms, and so forth⁵⁹. This technique has been 'reinvented' in the counting of gene fusions across many genomes as a predictor, for example, for protein–protein interactions or participation of proteins in common pathways⁶⁰. In both cases, practical implementations call for such steps as filtering of 'promiscuous' elements that are less predictive of common function or meaning⁶¹.

Literary linguistics and the genome

What might be called 'literary linguistics' includes pursuits ranging from stylistics to textual analysis to literary criticism. Although seemingly at opposite poles from the 'hard science' of molecular biology, these activities are at some level not so different from the increasingly hermeneutic role of the bioinformatician, insofar as

both are concerned with comparing texts, detecting subtle patterns and relationships, and elucidating theme and variation²⁵. Nor is textual criticism devoid of quantitative methods; concern with issues such as authorship attribution and authenticity has engendered an active discipline of statistical literary studies aided by computing^{62,63}.

The most pervasive theme in all such work is the study of word frequencies in texts, the mathematical analysis of which originates with the linguist G. K. Zipf, who first observed a power-law distribution relating a word's frequency of occurrence to the inverse of its position in the rank ordering of those frequencies⁶⁴ (Fig. 3). Mandelbrot elaborated on this insight, proposing a relationship between what has come to be known as Zipf's law and a presumed fractal nature of languages⁶⁵. Apparent instances of power-law behaviour have now been observed in many facets of molecular biology, including oligonucleotide frequencies⁶⁶, sizes of gene families⁶⁷ (including pseudogenes⁶⁸), distributions of protein⁶⁹ and RNA⁷⁰ folds, and even levels of gene expression⁷¹. As in the linguistic case, several explanations for these power-law behaviours have been proposed, including their mathematical relationship to scale-free networks⁷² such as might be expected in metabolic pathways⁷³ and protein interaction maps⁷⁴, and models for how they might arise in the evolution of protein families⁶⁹, all of which evince comparison to properties of words.

Textual criticism shares both goals and methods with bioinformatics. Species-specific distributions of oligonucleotides are among the signals (called style markers by linguists⁶²) that have been used in 'authorship attribution' of genome segments thought to arise by horizontal transmission between species (for example, pathogenicity islands in bacteria⁷⁵), and in checking the 'authenticity' of cloned sequences possibly contaminated by foreign material⁷⁶. Word frequencies and many other style markers have been analysed in literature using such tools as clustering⁷⁷, principal components analysis⁷⁸, neural networks⁷⁹, support vector machines⁸⁰ and genetic algorithms⁸¹, all of which are now being applied as well to 'transcript frequencies' inferred from microarray experiments. A recent review of these methods applied to gene expression comes full circle by using a clustering algorithm to group and classify articles on the topic based on word frequencies⁸², a foray into what has been termed bibliomics^{33,83}.

The complexity of human and biological-sequence languages at a lexical level has been compared explicitly by Trifonov and co-workers⁸⁴. Using metrics designed to detect the extent of 'overlapping codes', they suggest that sequence languages are more layered, with multiple signals reflecting, for example, different cellular processes, and thus more 'complex' insofar as the codes may constrain or interfere with one another⁸⁵. (Extreme examples are viral genomes with overlapping, frameshifted coding regions.) It should be noted, however, that human language is not 'single code' as suggested by Trifonov, but involves layering at multiple levels. An obvious illustration is poetry, where lexical and syntactic accommodations are often made for such overlaid constraints as rhyme scheme, metre and verse form, and even higher orders of metaphor, mood and theme — witness the virus-like economy of a *haiku*. Such superposition in languages is even treated formally, insofar as context-free languages are not closed under intersection and thus may be driven higher in the Chomsky hierarchy by layering¹⁰, a specific instance is the view of a pseudoknot as the intersection of two stem-loop structures⁴⁰.

A branch of textual criticism called stemmatics is concerned with the accuracy of texts, possibly ancient, that exist in multiple forms for reasons ranging from printers' errors to authorial revisions to fragmentary sources. For manuscripts copied many times by scribes, there has even been mathematical modelling of copying errors for purposes of estimating pairwise distances along a path from a common ancestor⁸⁶; biologically motivated algorithms have been enlisted in this cause to elucidate the provenance of Chaucer's *Canterbury Tales*⁸⁷. However, the very foundation of these algorithms in biological cladistics recapitulates older, similar methods from stemmatics and linguistics, as was already recognized a quarter-century ago⁸⁸.

One post-modern (and thus antiauthoritarian) school of textual criticism promotes the idea of a genetic text, a dynamic concept that encompasses all versions and even sources of a text through time⁸⁹, largely abandoning the concept of a 'main' version and thereby requiring new organizational paradigms and computational aids⁹⁰. The genetic text that is the genome surely presents similar challenges, and the many commonalities (as well as the instructive differences) between natural and biological languages may thus form the basis for sharing tools, techniques and ways of thinking about complex systems, on many different levels. □

doi:10.1038/nature01255

1. Aitchison, J. *Linguistics* (NTC/Contemporary Publishing, Chicago, 1999).
2. Chomsky, N. *Syntactic Structures* (Mouton, The Hague, 1957).
3. Jurafsky, D. & Martin, J. H. *Speech and Language Processing* (Prentice Hall, Upper Saddle River, NJ, 2000).
4. Brendel, V. & Busse, H. G. Genome structure described by formal languages. *Nucleic Acids Res.* **12**, 2561–2568 (1984).
5. Head, T. Formal language theory and DNA: an analysis of the generative capacity of specific recombinant behaviors. *Bull. Math. Biol.* **49**, 737–759 (1987).
6. Searls, D. B. in *Proc. 7th Natl Conf. Artif. Intell.* 386–391 (AAAI Press, Menlo Park, CA, 1988).
7. Searls, D. B. The linguistics of DNA. *Am. Sci.* **80**, 579–591 (1992).
8. Searls, D. B. in *Logic Programming: Proc. North Am. Conf.* (eds Lusk, E. & Overbeek, R.) 189–208 (MIT Press, Cambridge, MA, 1989).
9. Searls, D. B. in *Artificial Intelligence and Molecular Biology* Ch. 2 (ed. Hunter, L.) 47–120 (AAAI Press, Menlo Park, CA, 1993).
10. Searls, D. B. in *Mathematical Support for Molecular Biology* (eds Farach-Colton, M., Roberts, F. S., Vingron, M. & Waterman, M.) 117–140 (American Mathematical Society, Providence, RI, 1999).
11. Durbin, R., Krogh, A., Mitchison, G. & Eddy, S. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge Univ. Press, Cambridge, 1998).
12. Baldi, P. & Brunak, S. *Bioinformatics: The Machine Learning Approach* (MIT Press, Cambridge, MA, 2001).
13. Lyngso, R. B. & Pedersen, C. N. RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.* **7**, 409–427 (2000).
14. Joshi, A. in *Natural Language Processing: Psycholinguistic, Computational and Theoretical Perspectives* (eds Dowty, D., Karttunen, L. & Zwicky, A.) 206–250 (Chicago Univ. Press, New York, 1985).
15. Uemura, Y., Hasegawa, A., Kobayashi, S. & Yokomori, T. Tree-adjointing grammars for RNA structure prediction. *Theor. Comput. Sci.* **10**, 277–303 (1999).
16. Searls, D. B. String Variable Grammar: a logic grammar formalism for DNA sequences. *J. Logic Program.* **24**, 73–102 (1995).
17. Rivas, E. & Eddy, S. R. The language of RNA: a formal grammar that includes pseudoknots. *Bioinformatics* **16**, 334–340 (2000).
18. Shieber, S. Evidence against the context-freeness of natural language. *Linguist. Phil.* **8**, 333–343 (1985).
19. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. SMART, a simple modular architecture research tool: identification of signalling domains. *Proc. Natl Acad. Sci. USA* **95**, 5857–5864 (1998).
20. Westhead, D. R., Slidel, T. W., Flores, T. P. & Thornton, J. M. Protein structural topology: automated analysis and diagrammatic representation. *Protein Sci.* **8**, 897–904 (1999).
21. Abe, N. & Mamitsuka, H. Predicting protein secondary structure using stochastic tree grammars. *Machine Learn.* **29**, 275–301 (1997).
22. Przytycka, T., Srinivasan, R., & Rose, G. D. Recursive domains in proteins. *Protein Sci.* **11**, 409–417 (2002).
23. Jung, J. & Lee, B. Circularly permuted proteins in the protein structure database. *Protein Sci.* **10**, 1881–1886 (2001).
24. Hopcroft, J. E. & Ullman, J. D. *Introduction to Automata Theory, Languages, and Computation* (Addison-Wesley, Reading, MA, 1979).
25. Searls, D. B. Reading the book of life. *Bioinformatics* **17**, 579–580 (2001).
26. Dong, S. & Searls, D. B. Gene structure prediction by linguistic methods. *Genomics* **23**, 540–551 (1994).
27. Searls, D. B. Linguistic approaches to biological sequences. *Comput. Appl. Biosci.* **13**, 333–344 (1997).
28. Collado-Vides, J. A transformational-grammar approach to the study of the regulation of gene expression. *J. Theor. Biol.* **136**, 403–425 (1989).
29. Rosenbluth, D. A. et al. Syntactic recognition of regulatory regions in *Escherichia coli*. *Comput. Appl. Biosci.* **12**, 15–22 (1996).
30. Leung, S., Mellish, C. & Robertson, D. Basic Gene Grammars and DNA-ChartParser for language processing of *Escherichia coli* promoter DNA sequences. *Bioinformatics* **17**, 226–236 (2001).
31. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
32. Reese, M. G., Kulp, D., Tammana, H. & Haussler, D. Genie—gene finding in *Drosophila melanogaster*. *Genome Res.* **10**, 529–538 (2000).
33. Yandell, M. D. & Majoros, W. H. Genomics and natural language processing. *Nature Rev. Genet.* **3**, 601–610 (2002).
34. Sakakibara, Y. et al. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res.* **22**, 5112–5120 (1994).
35. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
36. Rivas, E. & Eddy, S. R. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2**, 8 (2001).
37. Knudsen, B. & Hein, J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics* **15**, 446–454 (1999).
38. Brown, M. P. Small subunit ribosomal RNA modeling using stochastic context-free grammars. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **8**, 57–66 (2000).
39. Holmes, I. & Rubin, G. M. Pairwise RNA structure comparison with stochastic context-free grammars. *Pac. Symp. Biocomput.* 163–174 (2002).
40. Brown M. & Wilson C. RNA pseudoknot modeling using intersections of stochastic context free grammars with applications to database search. *Pac. Symp. Biocomput.* 109–125 (1996).

41. Campbell, L. *Historical Linguistics: An Introduction* (MIT Press, Cambridge, MA, 1999).
42. Darwin, C. *The Descent of Man* (John Murray, London, 1871).
43. Dawkins, R. *The Selfish Gene* (Oxford Univ. Press, Oxford, 1976).
44. Nowak, M. A., Komarova, N. L. & Niyogi, P. Computational and evolutionary aspects of language. *Nature* **417**, 611–617 (2002).
45. Pennock, R. T. *Tower of Babel: The Evidence against the New Creationism* (Bradford/MIT Press, Cambridge, MA, 1999).
46. Cavalli-Sforza, L. L. *Genes, Peoples, and Languages* (North Point Press, New York, 2000).
47. Warnow T. Mathematical approaches to comparative linguistics. *Proc. Natl Acad. Sci. USA* **94**, 6585–6590 (1997).
48. Swadesh, M. Lexicostatistical dating of prehistoric ethnic contacts: with special reference to North American Indians and Eskimos. *Proc. Am. Phil. Soc.* **96**, 452–463 (1952).
49. Kruskal, J. B., Dyen, I. & Black, P. in *Lexicostatistics in Genetic Linguistics* (ed. Dyen, I.) 30–55 (Mouton, The Hague, 1973).
50. Mushegian, A. The minimal genome concept. *Curr. Opin. Genet. Dev.* **9**, 709–714 (1999).
51. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**, 33–36 (2000).
52. Snel, B., Bork P. & Huynen, M. A. Genome phylogeny based on gene content. *Nature Genet.* **21**, 108–110 (1999).
53. Tekai, F., Lazcano, A., & Dujon, B. The genomic tree as revealed from whole proteome comparisons. *Genome Res.* **9**, 550–557 (1999).
54. Lin, J. & Gerstein, M. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res.* **10**, 808–818 (2000).
55. Pellegrini, M. et al. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. USA* **96**, 4285–4288 (1999).
56. McWhorter, J. H. *The Power of Babel: A Natural History of Language* 128–129 (Freeman, New York, 2001).
57. Searls, D. B. From *Jabberwocky* to genome: Lewis Carroll and computational biology. *J. Comp. Biol.* **8**, 339–348 (2001).
58. Lupas, A. N., Ponting, C. P. & Russell, R. B. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* **134**, 191–203 (2001).
59. McKeown, K. R. & Radev, D. R. in *A Handbook of Natural Language Processing* (eds Dale, R., Moisl, H. & Somers, H.) 507–523 (Dekker, New York, 2000).
60. Marcotte, E. M. et al. Detecting protein function and protein-protein interactions from genome sequences. *Science* **285**, 751–753 (1999).
61. Smadja, F. Retrieving collocations from text: XTRACT. *Comput. Linguist.* **19**, 143–177 (1993).
62. Rudman, J. The state of authorship attribution studies: some problems and solutions. *Comput. Humanities* **31**, 351–365 (1998).
63. Barnbrook, G. *Language and Computers* (Edinburgh Univ. Press, Edinburgh, 1996).
64. Zipf, G. K. *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Boston, MA, 1949).
65. Mandelbrot, B. *The Fractal Geometry of Nature* (Freeman, San Francisco, 1983).
66. Mantegna, R. N. et al. Linguistic features of noncoding DNA sequences. *Phys. Rev. Lett.* **73**, 3169–3172 (1994).
67. Huynen, M. A. & van Nimwegen, E. The frequency distribution of gene family sizes in complete genomes. *Mol. Biol. Evol.* **15**, 583–589 (1998).
68. Harrison, P. M. & Gerstein, M. Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J. Mol. Biol.* **318**, 1155–1174 (2002).
69. Qian, J., Luscombe, N. M. & Gerstein, M. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J. Mol. Biol.* **313**, 673–681 (2001).
70. Schuster, P., Fontana, W., Stadler, P. F. & Hofacker, I. L. From sequences to shapes and back: a case study in RNA secondary structures. *Proc. R. Soc. Lond. B* **255**, 279–284 (1994).
71. Hoyle, D. C., Rattray, M., Jupp, R. & Brass, A. Making sense of microarray data distributions. *Bioinformatics* **18**, 576–584 (2002).
72. Rzhetsky, A. & Gomez, S. M. Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics* **17**, 988–996 (2001).
73. Jeong, H. et al. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
74. Park, J., Lappe, M. & Teichmann, S. A. Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.* **307**, 929–938 (2001).
75. Garcia-Valle, S., Romeu, A. & Palau, J. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.* **10**, 1719–1725 (2000).
76. White, O. et al. A quality control algorithm for DNA sequencing projects. *Nucleic Acids Res.* **21**, 3829–3838 (1993).
77. Hoover, D. I. Statistical stylistics and authorship attribution: an empirical investigation. *Lit. Linguist. Comput.* **16**, 421–444 (2001).
78. Binongo, J. N. G. & Smith, M. W. A. The application of principal component analysis to stylometry. *Lit. Linguist. Comput.* **14**, 445–466 (1999).
79. Hoorn, J. F., Frank, S. L., Kowalczyk, W. & van der Ham, F. Neural network identification of poets using letter sequences. *Lit. Linguist. Comput.* **14**, 311–338 (1999).
80. Leopold, E. & Kindermann, J. Text categorization with support vector machines. How to represent texts in input space? *Machine Learn.* **46**, 423–444 (2002).
81. Holmes, D. I. & Forsyth, R. S. The Federalist revisited: new directions in authorship attribution. *Lit. Linguist. Comput.* **10**, 111–127 (1995).
82. Altman, R. B. & Raychaudhuri, S. Whole-genome expression analysis: challenges beyond clustering. *Curr. Opin. Struct. Biol.* **11**, 340–347 (2001).
83. Searls D. B. Mining the bibliome. *Pharmacogenomics J.* **1**, 88–89 (2001).
84. Popov, O., Segal, D. M. & Trifonov, E. N. Linguistic complexity of protein sequences as compared to texts of human languages. *Biosystems* **38**, 65–74 (1996).
85. Trifonov, E. N. Interfering contexts of regulatory sequence elements. *Comput. Appl. Biosci.* **12**, 423–429 (1996).
86. Spenser, M. & Howe, C. Estimating distances between manuscripts based on copying errors. *Lit. Linguist. Comput.* **16**, 467–484 (2001).
87. Barbrook, A. C., Howe, C. J., Blake, N. & Robinson, P. The phylogeny of the *Canterbury Tales*. *Nature* **394**, 839 (1998).
88. Platnick, N. I. & Cameron, H. D. Cladistic methods in textual, linguistic, and phylogenetic analysis. *Syst. Zool.* **26**, 380–385 (1977).
89. Tanselle, G. T. *Literature and Artifacts* (Bibliographical Society of the University of Virginia, Charlottesville, VA, 1998).
90. Ferrer, D. Hypertextual representation of literary working papers. *Lit. Linguist. Comput.* **10**, 143–145 (1995).

Acknowledgements

I thank P. Agarwal, A. Lupas, N. Odendahl and K. Rice for helpful comments on the manuscript.

The structure of the protein universe and genome evolution

Eugene V. Koonin, Yuri I. Wolf & Georgy P. Karev

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA (e-mail: koonin@ncbi.nlm.nih.gov)

Despite the practically unlimited number of possible protein sequences, the number of basic shapes in which proteins fold seems not only to be finite, but also to be relatively small, with probably no more than 10,000 folds in existence. Moreover, the distribution of proteins among these folds is highly non-homogeneous — some folds and superfamilies are extremely abundant, but most are rare. Protein folds and families encoded in diverse genomes show similar size distributions with notable mathematical properties, which also extend to the number of connections between domains in multidomain proteins. All these distributions follow asymptotic power laws, such as have been identified in a wide variety of biological and physical systems, and which are typically associated with scale-free networks. These findings suggest that genome evolution is driven by extremely general mechanisms based on the preferential attachment principle.

The distribution of matter and energy in the Universe provides cosmologists with the principal source of information on the evolution of our planet, including its earliest stages. In particular, the discovery of the uniformly distributed background microwave radiation is the main proof of the Big Bang model of the Universe's origin (for example, see refs 1, 2). In a somewhat loose but perhaps appropriate analogy, structural biologists often speak of the 'protein universe', meaning the totality of all possible proteins^{1–3}. The total number of possible protein sequences (that is, the size of the protein universe) is, for all practical purposes, infinite. Assuming an average protein length of 200 amino acids, there can be 20^{200} different protein sequences, a number that is much greater than, for example, the number of electrons in our (physical) Universe.

Our current theoretical understanding of protein folding is insufficient to estimate the total possible number of protein structures, but it too is likely to be vast. Obviously, only a minuscule fraction of the potential sequence space is populated by real protein sequences, but the number of unique sequences encoded in actual genomes is likely to be substantial. For example, assuming there are 10 million species on Earth and the genome of each species consists of 5,000 genes (an intermediate number between prokaryotes and eukaryotes), there are 5×10^{10} unique protein sequences. Although this quantity is negligible compared to the vast sequence space, it still is several orders of magnitude greater than that contained in today's databases. A question of fundamental and practical interest is how these sequences are distributed in the sequence and structure spaces.

The protein universe is an abstraction, however useful. In reality, all proteins are, of course, encoded in genes, which belong to particular genomes. Quantitative and qualitative analysis of the projections of the protein universe on genomes from a diverse range of organisms might reveal important aspects of the evolution of both genomes and proteins.

Distribution of protein families and protein folds

That the population of the protein universe is not distributed randomly is obvious from the existence of homologous genes and proteins. However, to extract any useful

information from this distribution, it needs to be explored in quantitative detail, which can be done only within the framework of a hierarchical taxonomy of proteins. Margaret Dayhoff's group introduced the notions of protein family and superfamily in the 1960s as part of their effort to understand protein evolution and simultaneously create a well-organized protein database^{4,5} (later known as the Protein Identification Resource). A family was defined as a group of (closely) related sequences, and superfamilies encompassed two or more related families.

By the mid-1990s, a more elaborate and coherent taxonomy of protein domains had been developed, largely through the efforts of Murzin and colleagues, who constructed the SCOP classification of protein structures, and Thornton and colleagues, who produced the CATH database dedicated to the same goal^{1,6–9}. The top levels of the hierarchy are defined by the three-dimensional structure, whereas lower taxa are identified on the basis of sequence similarity and functional considerations (Table 1). Exact criteria of topological similarity, which is necessary and sufficient to assign two protein structures to the same fold, or the level of sequence similarity that defines a superfamily or a family, have yet to be determined in full. Nevertheless, there is a wide agreement both on the general principles of classification and on the taxonomic assignments of most proteins^{6,8,10,11}.

At this point, it is important to introduce the fundamental notion of protein domain, which is the foundation of at least the top levels of the protein taxonomy. In structural biology, a domain is defined as a distinct, compact and stable protein structural unit that folds independently of other such units¹². Often, however, domains are characterized differently — as distinct regions of protein sequence that are highly conserved in evolution. As the hierarchy of protein classification evolved into a combination of structural- and sequence-based approaches, the notions of structural and 'homology' domains also tended to blend into one concept. The salient features of structural domains (that is, independent folding and stability) conduce them to become distinct evolutionary units, which exist as stand-alone proteins or as parts of various domain architectures in multidomain proteins. There are exceptions to this generalization, one

Table 1 Hierarchical classification of proteins

Category	Example	Definition, criteria or main features
Structural class	α/β	Overall composition of structural elements. No evolutionary relationship.
Fold	TIM barrel	Topology of the folded protein backbone. Monophyletic origin?
Superfamily	Aldolase	Recognizable sequence similarity (at least a conserved motif); conservation of basic biochemical properties. Monophyletic origin.
Family	Class I aldolase	Significant sequence similarity; conservation of biochemical activity.
Group of orthologues (COG)	2-keto-3-deoxy-6-phosphogluconate aldolase	Orthologous relationships within the given set of species; conservation of the biochemical activity and, most often, also the biological function.
Lineage-specific expansion	PA3131 and PA3181 in <i>Pseudomonas aeruginosa</i>	Paralogues originating from a lineage-specific duplication; possible functional specialization.

being where two structural domains comprise a seemingly inseparable evolutionary unit (a 'homology domain'). But whenever this situation is observed, stand-alone versions or new multidomain architectures of the respective domains are usually discovered eventually; this is supported by numerous observations made in the context of recent genome analyses (for example, see refs 13, 14).

There is no doubt that protein families and superfamilies are monophyletic, that is, they derive from a common ancestor. In contrast, monophyly of protein folds, as opposed to folds originating by convergence from unrelated ancestors, remains an issue of debate. It seems that, for most folds (with the possible exception of some of the most diverse 'superfolds'), similarity goes beyond the topology of the protein backbone. Often, the basic physicochemical interactions and the associated structural and sequence motifs are conserved throughout a fold (for example, the P-loop in the eponymous ATP/GTPase fold^{15,16}), or even across fold boundaries (for example, the phosphate-binding loop in Rossmann-type nucleotide-binding domains¹⁷). Perhaps more important, on numerous occasions, the same activity and/or function is performed by two or more unrelated folds in different organisms or in different cellular systems in the same organism^{18,19}. Taken together, these observations seem to argue against convergence as the prevalent force in the evolution of protein folds and suggest that most, if not all, protein folds are monophyletic. However, the possibility of multiple, convergent origins still might be considered for some common folds with a relatively simple, symmetric topology, such as TIM barrels (named after the structure of the glycolytic enzyme triosephosphate isomerase) or β -propellers.

Protein families consist of related 'individuals', each of which is a set of orthologues, or proteins related by vertical descent (according to the classification of homologues proposed by Walter Fitch^{20,21}). Clusters of orthologous groups of proteins (COGs) typically occupy a unique functional niche, which remains the same in different, even phylogenetically distant organisms, except for lineage-specific expansions of proteins within a COG^{22,23}. These expansions that result from relatively recent duplications are prominent in genomes, particularly in eukaryotes^{24–26}. On many occasions, there is a plausible connection between the lineage-specific proliferation of a particular family and specific adaptations characteristic of the given group of organisms. The relationships between distinct COGs within a family (as well as between families within a superfamily and, most likely, between superfamilies within a fold) represent paralogy, that is, origin from an ancestral duplication^{20,21,27}. Paralogous COGs within a family tend to have different biological functions, although, in many cases, they have identical or similar biochemical activities.

Early sequence and structure databases were severely biased, primarily because of overrepresentation of sequences of well-characterized proteins and gross under-representation of uncharacterized ones. Growth of the databases, especially with the advent of high-throughput genome sequencing, eliminated much of this sampling

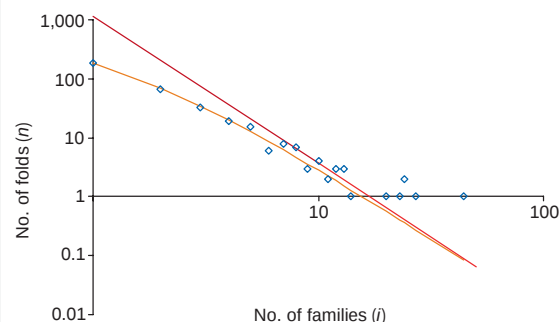


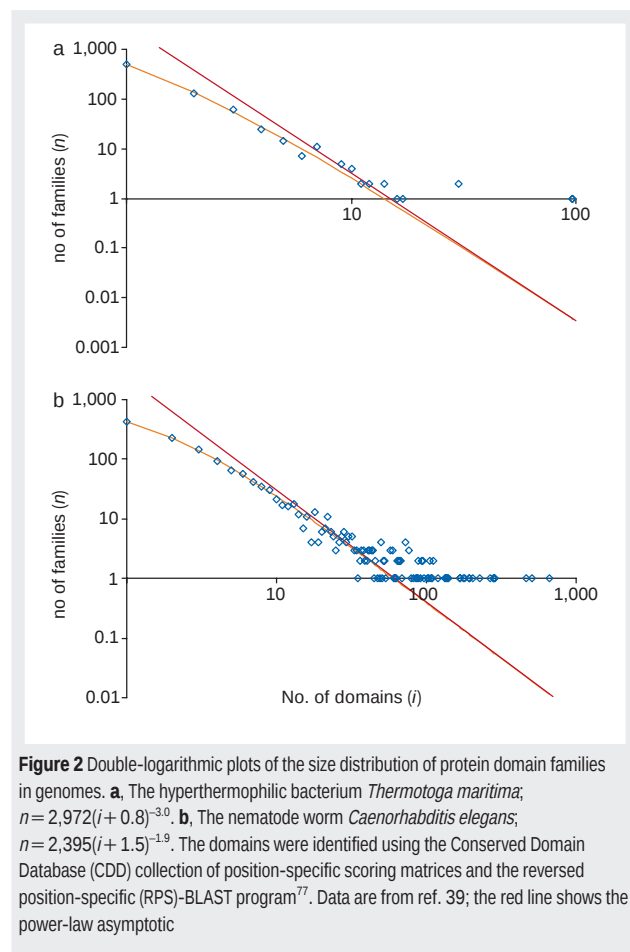
Figure 1 Double-logarithmic plot of the distribution of protein folds by the number of families. The sequences from the Structural Classification of Proteins (SCOP) 1.39 database were analysed as described in ref. 71. The best fit is defined by the equation $n = 1,165(i + 1.1)^{-2.5}$. The red line shows the power-law asymptotic.

bias. By mid-1990, it became clear that the distribution of protein domains among folds, superfamilies and families was extremely uneven — most taxa consisted of a small number of members and only a few were highly abundant. Rigorous application of sampling theory ruled out sampling bias as the principal contribution to the observed distribution^{28,29}.

As the sampled fraction of the protein universe increased, more reliable estimates of the overall variety of proteins became feasible. In contrast to the earlier assessments, which relied largely on the rate of discovery of new protein families^{4,30,31}, these studies used the observed distributions of families among folds to extrapolate the total numbers, taking the sampling process into account. Depending on the assumptions and methods used, the estimates of the total number of existing protein folds produced by different researchers varied substantially, from ~650 to ~10,000 (refs 29, 32–37). But examination of the distribution of folds by the number of protein families (Fig. 1) indicates that, in one sense, the discrepancy between these estimates might be of little consequence. This distribution contains a small number of folds with a large number of families (mostly well-known superfolds, such as P-loop NTPases, the Rossmann fold or TIM barrels) and an increasing number of folds that consist of a small number of families. By far the largest size class consists of the 'unifolds'³⁷, each including one family, often just one COG. Thus, it seems certain that the great majority of protein families belong to ~1,000 common folds. What is still in dispute is the number of unifolds that encompass the rest of the proteins. Approximately one half of the common folds are currently represented by at least one experimentally determined structure, which means that coarse-grain mapping of the protein universe is already at an advanced stage.

Power laws and models of genome evolution

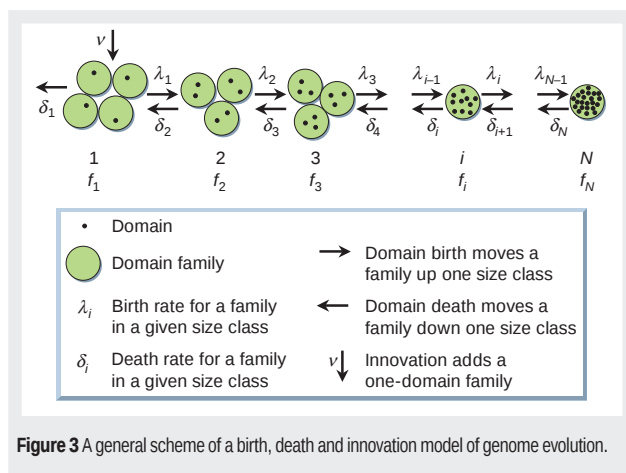
Mathematically, the distribution of protein folds by the number of constituent families has been approximated by a power law, that is, $f(i) \sim i^{-\gamma}$ where $f(i)$ is the frequency of folds that include exactly i families and γ is a parameter that typically assumes values between 1 and 3 (ref. 34). More precisely, the fold–family distribution fits the so-called generalized Pareto function $f(i) \sim (i + a)^{-\gamma}$, where a is an additional parameter, with the power law fitting asymptotically with the increase of i (Fig. 1). Remarkably, the same function, up to the parameters, fits the distribution of protein domain families by the number of members in each analysed genome, as recently shown by Kuznetsov³⁸ and by ourselves^{39,40} (Fig. 2). These distributions, along with the distributions of other genome-associated quantities (for example, the number of pseudogenes per gene family), have been previously approximated with power laws, first in the pioneering work of Huynen and Van Nimwegen⁴¹ and subsequently in detailed studies by Gerstein and colleagues^{42–44}.



As demonstrated by Barabasi and colleagues and by several other researchers, power laws describe the distribution of various quantities in numerous biological, physical and social contexts; such distributions can seem to be fundamentally different (for example, the number of links between documents in the Internet, the population of towns and the number of reactions in which a given metabolite is involved)^{45–50}. Zipf's law, which describes the frequency distribution of words in texts⁵¹, and the Pareto principle, describing the distribution of people by wealth⁵², are in this category. These distributions have specific mathematical properties related to those of so-called scale-free networks, that is, networks in which the frequency distribution of node degrees (the number of nodes to which a given node is connected) follows a power law^{47,48}. In particular, the network of metabolic reactions in any organism is a scale-free network with a distinct hierarchical structure^{50,53} and protein–protein interaction networks have similar properties⁵⁴.

The wide spread of power distributions and scale-free networks in nature and society suggests that similar laws might govern evolution in a variety of diverse systems. The general pattern of network evolution that ensures scale-free behaviour is preferential attachment, where the probability of a node acquiring a new connection is proportional to the degree (the number of connections) of that node. Metaphorically, this can be described as a situation in which 'the rich get richer' or, from a selectionist perspective, 'the fit get fitter'⁴⁷.

Returning to protein domains, there seems to be at least three (not necessarily exclusive) ways to explain the emergence of power laws and related highly skewed distributions of the fold and family sizes in the protein universe and in individual genomes. The 'designability' hypothesis, favoured by some structural biologists, postulates that certain folds serve as attractors in the space of protein structures because of their topological properties (for example, the highly



abundant TIM-barrel fold is a uniquely symmetrical construction). As a result, many unrelated sequences tend to adopt the same few folds. Interestingly, the simulated designability distributions analysed by Wingreen and colleagues^{55,56} appear to be similar to the empirical distributions of domain family sizes described by Gerstein and co-workers^{42,44}, Kuznetsov³⁸ and ourselves³⁹. However, given the above argument against a convergent origin of most folds, designability does not seem to be a likely general explanation for the observed preferential attachment or, more precisely, preferential proliferation of domains in protein evolution.

A straightforward selectionist interpretation holds that certain biochemical activities (for example, nucleoside 5'-triphosphate hydrolysis), being particularly common and important in cellular biochemistry, are in greater demand than other, highly specialized ones, which leads to preferential proliferation of the respective protein families. Again, the weakness of this argument is that the same activity is often embodied in two or more distinct domains, which tend to differ substantially in abundance¹⁸.

Finally, domain birth and death models developed by Gerstein and co-workers⁴², Rzhetsky and Gomez⁵⁷ and ourselves³⁹, which originate from the classic analysis of Yule⁵⁸, completely disregard the protein identity, but give rise to equilibrium distributions of domain family sizes that show an excellent fit to the observed ones. These models typically include the elementary processes of family growth via domain birth (duplication), domain death as a result of inactivation and loss, and innovation or emergence of a new family (for example, through extensive modification of a member of an existing family, horizontal gene transfer or even origin of a new protein from non-coding sequence) (Fig. 3).

We recently explored the behaviour of these birth, death and innovation models (BDIMs) in detail, both analytically and by computer simulation; this analysis seems to lead to non-trivial conclusions on genome evolution³⁹. First, it was shown that, using BDIMs, an equilibrium distribution of domain family sizes is reached exponentially fast during evolution from any initial conditions. Specifically, $|f_i(t) - f_i| \sim e^{-kt}$, where $f_i(t)$ is the frequency of a given family at time t and f_i is the equilibrium frequency. Thus, any perturbation in genome evolution, which involves changes in the parameters of birth, death or innovation, rapidly relaxes to a new stationary state. Accordingly, the mode of evolution depicted by BDIMs is most compatible with the punctuated equilibrium notion of genome evolution⁵⁹. By this model, long periods of stasis are punctuated by relatively brief bursts of evolutionary activity, which involve rapid proliferation and elimination of gene families as well as 'invention' and acquisition of new ones.

Second, BDIMs result in different shapes of equilibrium distributions of family sizes depending on how precisely the birth rate is balanced by the death rate. The power law appears as an asymptotic in a certain, specific subclass of BDIM, in which the death rate

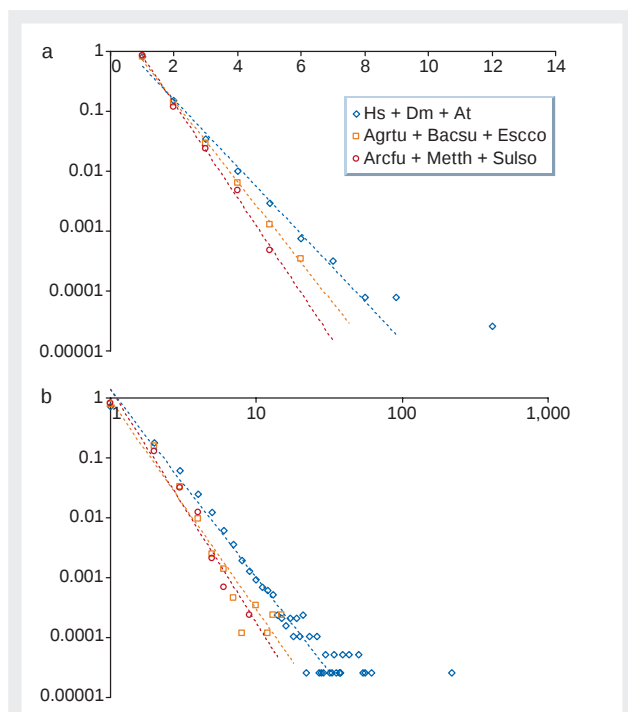


Figure 4 Distributions of the number of domains in proteins from the three primary kingdoms of life. **a**, Repeats of the same domain in a single polypeptide excluded. The plot is in semi-logarithmic scale. **b**, Repeats of the same domain in a single polypeptide included. The plot is in double logarithmic scale. The data and methods used for generating this plot were the same as in Fig. 2. Eukaryotes: Hs, *Homo sapiens*; Dm, *Drosophila melanogaster*; At, *Arabidopsis thaliana*. Bacteria: Agrtu, *Agrobacterium tumefaciens*; Bacsu, *Bacillus subtilis*; Escco, *Escherichia coli*. Archaea: Arcfu, *Archaeoglobus fulgidus*; Metth, *Methanothermobacter thermoautotrophicus*; Sulso, *Sulfolobus solfataricus*.

approaches the birth rate for large families, but is considerably greater than the birth rate for small families. These models accurately describe the distributions of domain family size for all analysed genomes, whereas straightforward approximation with a power law does not fit the data nearly as well (Fig. 2).

Finally, analysis of BDIMs shows that the innovation rate, which is required to offset the stochastic loss of low-copy families, has to be relatively high and, at least in small, prokaryotic genomes, comparable to the overall intra-genomic duplication (birth) rate. This supports, from a somewhat unexpected angle, the key role of horizontal gene transfer in prokaryotic evolution that has been suggested by numerous observations made during genome comparisons^{60–64}.

The evolutionary models described here ignore completely the individuality of gene families and the selective forces that make some of them expendable and others indispensable. Despite this obvious over-simplification, BDIMs accurately reproduce the observed family size distributions, suggesting that genome evolution might be largely a stochastic process, which is only modulated by selection.

Paradoxes of multidomain networks

Protein domains often combine to form multidomain architectures. Analysis of such architectures can be extremely helpful for predicting functions of uncharacterized domains and proteins in a ‘guilt by association’ approach (also called the ‘Rosetta Stone’ principle), which is based on the assumption that physical fusion of two domains implies a functional link^{65–68}. Indeed, multidomain proteins have critical roles in all living cells, as they provide effective links between different functional systems. Because of this ability, complex multidomain architectures are particularly characteristic of various signalling systems.

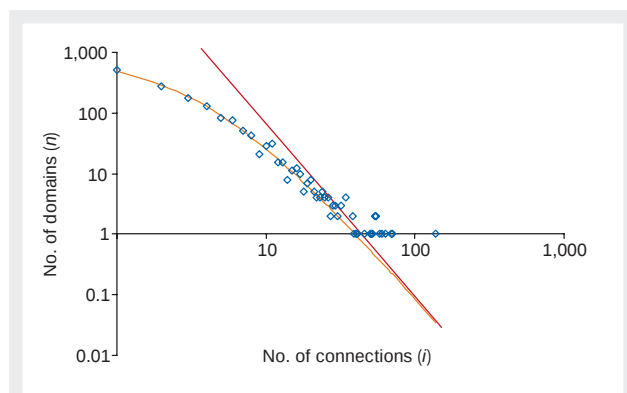


Figure 5 Double-logarithmic plot of the distribution of protein domains by the number of links in multidomain proteins. The number of links is the number of different domains with which the given domain combines in multidomain proteins. The combined data from seven analysed bacterial genomes, three archaeal genomes and six eukaryotic genomes were the same as in Fig. 2, except that several domains that showed artificially high numbers of connections because of their biased amino acid composition were removed manually. The best fit is given by the equation $n = 46,815(j + 4.0)^{-2.9}$.

There seems to be a connection between the propensity of protein domains to form multidomain architectures and the organismic complexity. Specifically, in many orthologous sets of eukaryotic proteins, such as chromatin-associated transcription factors, a distinct trend, which we dubbed ‘domain accretion’, can be traced towards increased complexity of domain architectures in more complex organisms⁶⁹. Because proteins form complex networks, even a modest increase in the number of domains in interacting partners could translate into numerous new interactions, which probably contributes to the solution of the apparent paradox of ‘too few’ genes in complex organisms⁷⁰.

Given the involvement of multidomain proteins in a variety of cellular functions, we might expect that natural selection should favour their formation to the extent that multidomain architectures would be over-represented with respect to single-domain proteins, especially in complex eukaryotes. However, quantitative analysis does not seem to support this conclusion. Instead, the distribution of proteins by the number of different domains (with multiple occurrences of the same domain in a given protein excluded from the analysis) shows an excellent fit to an exponent⁷¹ (Fig. 4a). This type of distribution is compatible with a random recombination (joining and breaking) model of evolution of multidomain architectures.

Notably, however, the slopes of the curves in Fig. 4a differ significantly for archaea, bacteria and eukaryotes, indicating that the fraction of multidomain proteins or, in terms of the random model, the likelihood of domain joining increases in the order: archaea < bacteria < eukaryotes. The under-representation of multidomain proteins in archaea compared to the other two primary kingdoms of life might be related to the low stability of large proteins in the hyperthermophilic habitats of most archaeal species. The excess of multidomain proteins in eukaryotes is not unexpected given the observations on domain accretion; furthermore, the right tail of the eukaryotic distribution shows a deviation from the exponent caused by the presence of several proteins with a large number of domains (Fig. 4a). When repeats of the same domain in a single polypeptide chain are added to the mix, the distribution changes and is best approximated by a generalized Pareto function (Fig. 4b). In light of the above, this finding does not seem unexpected: evolution of repeats is likely to follow a BDIM scenario, with tandem duplication and elimination as the main underlying processes, rather than a random joining–breaking model, which seems to apply to combinations of different domains.

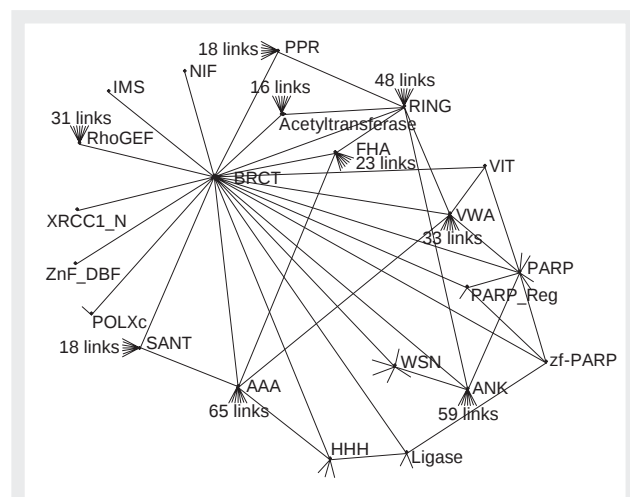


Figure 6 A fragment of the network of multidomain connections. All the connections of the BRCT (BRCA1 C-terminal) domain and those between its partners are shown; the number of outgoing connections is also indicated for all domains other than BRCT.

The above analysis does not tell us anything about the propensity of individual domains to form multidomain architectures, and these propensities differ widely. In an already familiar pattern, the distribution of the number of multidomain architectures in which a domain is involved roughly follows a power law, as demonstrated by Wuchty⁷² and by Teichmann and co-workers⁷³. More precisely, this distribution is described by a generalized Pareto function that we have already encountered in other contexts (Fig. 5). Thus, a small number of domains are hubs of multidomain connections that hold together cellular interaction networks. Although evolution of multidomain proteins containing different domains seems to occur primarily via random processes of joining and breaking (Fig. 4a), the fit (to form functionally advantageous multidomain architectures) still get fitter.

The network of multidomain connections for a moderately linked hub, the BRCT (BRCA1 C-terminal) domain, which is an important adaptor in eukaryotic cell-cycle checkpoints and DNA repair^{74,75}, is shown in Fig. 6. Notably, some of the domains linked to BRCT, such as RING (involved in ubiquitin-dependent cascades) and FHA (implicated in various signal-transduction pathways) are important hubs themselves. Table 2 shows the top multidomain connectors for bacteria, archaea and eukaryotes. Remarkably, the lists for the two prokaryotic kingdoms have five domains in common, whereas the eukaryotic list is completely different. Not unexpectedly, however, all three sets are dominated by domains that are involved in various forms of signal transduction and regulation of enzymatic activity.

Perspectives

The protein universe is extremely unevenly populated, with most proteins concentrated in a relatively small number of major clusters, the common folds and superfolds. This highly skewed distribution of proteins among folds should enable structural genomics research programmes to complete a preliminary tour of the most important part of the protein universe within the next few years⁷⁶, although many rare folds are likely to remain uncharacterized for much longer.

Projection of the structure of the protein universe on genomes and quantitative analysis of the outcome seems to result in some unexpected insights into general principles of genome evolution. Remarkably, the size distributions of folds for the explored part of the protein universe and of domain families for all analysed genomes, as well as the distribution of the number of domain connections in multidomain architectures, are all described by the same type of mathematical functions, in which the power law appears as an asymptotic. This suggests that extremely general mechanisms of

Table 2 Top multidomain connections in eukaryotes, bacteria and archaea

Eukaryotes		Bacteria		Archaea	
NC	Domain	NC	Domain	NC	Domain
133	Ser/Thr protein kinase	28	AAA ATPase	18	AAA ATPase
69	PH	23	Receiver domain	9	His-kinase-type ATPase
63	Ankyrin repeat	20	His-kinase-type ATPase	9	CBS
60	RRM	18	GAF	9	D-Ala-D-Ala ligase
58	Immunoglobulin	15	CBS	9	4Fe-4S ferredoxin
55	PHD finger	15	PAS	8	DEXD helicase
55	PDZ	14	His-kinase phosphoacceptor	8	PAS
54	SH3	14	HAMP	8	PAC
54	RING finger	14	FMO-like	8	FMO-like
52	C2	14	ACT	7	Receiver

Abbreviations and acronyms: NC, Number of connections; PH, pleckstrin homology domain; RRM, RNA-recognition motif; PHD, plant homeodomain (-associated finger); PDZ, postsynaptic density protein-95/discs large/zo-1 domain; SH3 Src homology 3 domain; FMO, flavin-dependent monooxygenase; GAF, CBS, PAS, PAC, HAMP, ACT are small-molecule-binding domains involved in various forms of signal transduction and allosteric regulation of enzymatic activity.

evolution, apparently based on the preferential attachment (proliferation) principle, are at work in all these contexts.

With respect to domain families, these principles have already been detailed in plausible, even if oversimplified, models of genome evolution based on the elementary processes of birth, death and innovation. Similar models could potentially be developed for other situations, such as the connections between domains in multidomain networks, as well as networks of protein–protein interactions and metabolic reactions. Evolutionary modelling certainly needs to be made more realistic by including additional parameters, particularly those associated with purifying and positive selection. It seems reasonable to hope that further quantitative analysis of the structure of the protein universe and its projections on diverse genomes ushers a qualitatively new understanding of the evolution of life in a not so remote future. □

doi:10.1038/nature01256

- Holm, L. & Sander, C. Mapping the protein universe. *Science* **273**, 595–603 (1996).
- Zhang, C. & DeLisi, C. Protein folds: molecular systematics in three dimensions. *Cell. Mol. Life Sci.* **58**, 72–79 (2001).
- Rost, B. Did evolution leap to create the protein universe? *Curr. Opin. Struct. Biol.* **12**, 409–416 (2002).
- Dayhoff, M. O. The origin and evolution of protein superfamilies. *Fed. Proc.* **35**, 2132–2138 (1976).
- Dayhoff, M. O., Barker, W. C. & Hunt, L. T. Establishing homologies in protein sequences. *Methods Enzymol.* **91**, 524–545 (1983).
- Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
- Murzin, A. G. Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.* **6**, 386–394 (1996).
- Orengo, C. A. *et al.* CATH—a hierarchical classification of protein domain structures. *Structure* **5**, 1093–1108 (1997).
- Todd, A. E., Orengo, C. A. & Thornton, J. M. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143 (2001).
- Lo Conte, L., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* **30**, 264–267 (2002).
- Orengo, C. A. *et al.* The CATH protein family database: a resource for structural and functional annotation of genomes. *Proteomics* **2**, 11–21 (2002).
- Branden, C. & Tooze, J. *Introduction to Protein Structure* (Garland Publishing, New York, 1999).
- Anantharaman, V., Koonin, E. V. & Aravind, L. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res.* **30**, 1427–1464 (2002).
- Anantharaman, V., Koonin, E. V. & Aravind, L. Regulatory potential, phylectic distribution and evolution of ancient, intracellular small-molecule-binding domains. *J. Mol. Biol.* **307**, 1271–1292 (2001).
- Saraste, M., Sibbald, P. R. & Wittinghofer, A. The P-loop—a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* **15**, 430–434 (1990).
- Koonin, E. V. A superfamily of ATPases with diverse functions containing either classical or deviant ATP-binding motif. *J. Mol. Biol.* **229**, 1165–1174 (1993).
- Aravind, L., Mazumder, R., Vasudevan, S. & Koonin, E. V. Trends in protein evolution inferred from sequence and structure analysis. *Curr. Opin. Struct. Biol.* **12**, 392–399 (2002).
- Galperin, M. Y., Walker, D. R. & Koonin, E. V. Analogous enzymes: independent inventions in enzyme evolution. *Genome Res.* **8**, 779–790 (1998).
- Martin, A. C. *et al.* Protein folds and functions. *Structure* **6**, 875–884 (1998).
- Fitch, W. M. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99–113 (1970).

- ## Acknowledgements

We thank A. Panchenko and S. He (NCBI) for help with the use of the Conserved Domain Database, and A. Rzhetsky and V. Kuznetsov for helpful discussions.

Engineered gene circuits

Jeff Hasty*, David McMillen† & J. J. Collins*

*Department of Bioengineering, University of California San Diego, La Jolla, California 92093, USA (e-mail: hasty@ucsd.edu)

†Center for BioDynamics and Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215, USA

A central focus of postgenomic research will be to understand how cellular phenomena arise from the connectivity of genes and proteins. This connectivity generates molecular network diagrams that resemble complex electrical circuits, and a systematic understanding will require the development of a mathematical framework for describing the circuitry. From an engineering perspective, the natural path towards such a framework is the construction and analysis of the underlying submodules that constitute the network. Recent experimental advances in both sequencing and genetic engineering have made this approach feasible through the design and implementation of synthetic gene networks amenable to mathematical modelling and quantitative analysis. These developments have signalled the emergence of a gene circuit discipline, which provides a framework for predicting and evaluating the dynamics of cellular processes. Synthetic gene networks will also lead to new logical forms of cellular control, which could have important applications in functional genomics, nanotechnology, and gene and cell therapy.

It has been over 40 years since Monod and Jacob boldly predicted that such fundamental cellular processes as differentiation and protein regulation are accomplished through signalling pathways resident at the level of the gene¹. This prediction laid the foundation for the ensuing progress in describing the essential regulatory mechanisms in many specific genetic systems. With the development of the field of nonlinear dynamics and the concurrent advent of significant computing power, mathematical models describing gene regulation began to appear regularly in the 1970s^{2–9}. Implicit in these studies was the realization that the ‘wiring’ of naturally occurring gene regulatory networks would be too complex for qualitative descriptions devoid of mathematics. Although this realization proved to be ahead of its time, owing mainly to the lack of experimentally deduced regulatory pathways in the ‘pre-genomic’ era, recent experimental advances have reignited interest in the development of circuit analysis techniques for describing complex gene networks.

The concept of designed gene circuits has motivated researchers to draw direct analogies with established techniques in electrical engineering^{10,11}. As with the construction of electrical circuits, the gene circuit approach^{12–24} uses mathematical and computational tools in the analysis of a proposed circuit diagram, while novel experimental techniques are used to construct the networks according to the model blueprint. So far, the qualitative agreement between model and experiment in a series of studies^{13–15,22,23} has supported the notion of such an engineering-based methodology (for a detailed discussion of the various mathematical modelling techniques and their particular applications, see refs 17 and 19). The power of this approach is that it can be used to study simplified systems to gain insight into the general ‘modules’ of gene regulation^{25–27}. These modules include subnetworks that act as switches or oscillators, as well as networks that act to communicate across a population of cells. This ability to engineer gene networks offers the prospect of extracting carefully chosen subsystems from natural organisms, and focusing both modelling and experimental effort on determination of the behaviour of the subsystems in isolation. Furthermore, there is the possibility of using the insights thus obtained to create genetic ‘control systems’, designed to

correct faulty cellular mechanisms, or to generate entirely new modes of behaviour.

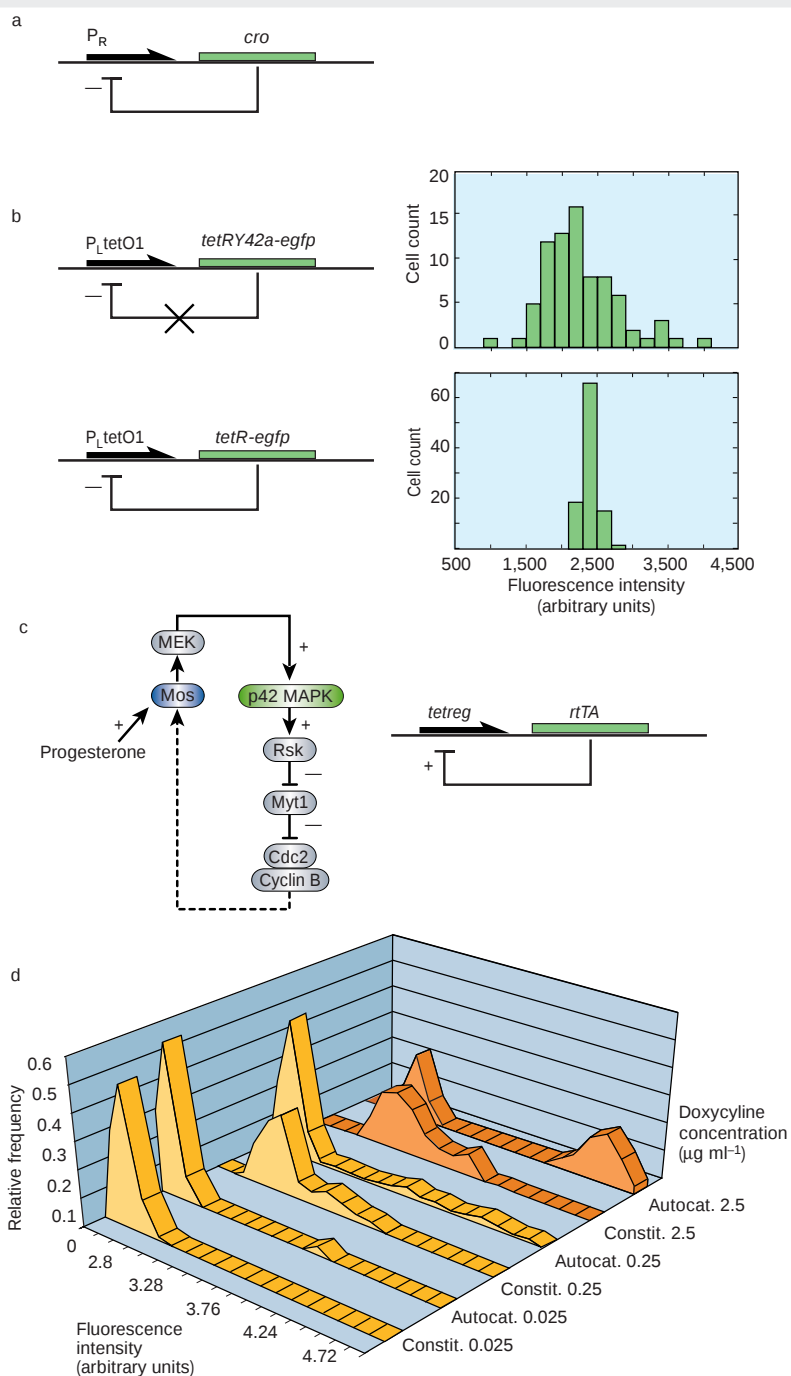
Autoregulatory systems

Feedback loops are an important concept in engineering control systems. In the context of gene regulation, feedback occurs through autoregulation, wherein a protein modifies, directly or indirectly, its own rate of production. Whether such interactions embody positive or negative feedback depends on the details of the network dynamics. Understanding the nature of such feedback loops in biological networks is a key step in the attempt to formulate a gene circuit discipline^{5–9,28–30}.

Seminal work in the modelling of gene networks³ focused on the stability properties of networks dominated by positive versus negative feedback. Stability refers to the tendency of a system to remain close to a steady state (a state in which production and decay rates are balanced) despite the influence of perturbations. A central result of this work was that genes regulated by negative feedback should be more stable than either unregulated genes or those regulated by positive feedback (an example of a naturally occurring negative feedback system is given in Fig. 1a). In the past few years, synthetic gene networks have been engineered to test the first portion of this prediction, comparing the behaviour of a simplified gene regulatory network based on negative feedback to the behaviour of the equivalent unregulated network¹⁵.

The experiments used a promoter that is shut off by the tetracycline repressor protein (TetR) to control the production of TetR, and compared this network to an unregulated system (Fig. 1b). Using a fusion of green fluorescent protein (GFP) to the TetR protein allowed observation of the state of the network (that is, the number of TetR molecules present in the cell) through fluorescence microscopy. Sampling multiple cells yielded a distribution of fluorescence intensities, and the stability of the steady state was evaluated by the width of these distributions. Because the degree of stability is inversely proportional to the width of the distribution, narrower distributions imply greater stability. Mathematical modelling indicated, as had the earlier analysis, that the negative feedback network should be more stable than the unregulated network, and the experimental results confirmed this prediction.

Figure 1 Autoregulatory systems. **a**, In this natural negative feedback system from the bacteriophage λ , the promoter P_R controls the expression of the Cro protein, which represses P_R . **b**, The synthetic negative feedback system¹⁵ uses the promoter P_{tetO1} to control the expression of TetR-EGFP, a fusion of the tetracycline repressor (TetR) and the enhanced green fluorescent protein (EGFP). Negative feedback arises because TetR represses transcription from P_{tetO1} . Replacing TetR with TetRY42A eliminates the feedback, producing an unregulated system. The distribution of observed expression states for the unregulated system (upper, right) is about three times wider than the distribution for the negative feedback system (lower, right), demonstrating improved stability with negative feedback. (Distributions redrawn from ref. 15.) **c** The left panel shows a natural positive feedback system, the Mos-MEK-p42 MAPK cascade, which controls part of the maturation process in *Xenopus* oocytes. Progesterone stimulates the production of the Mos protein, which indirectly activates p42 MAPK (mitogen-activated protein kinase). p42 MAPK activation, in turn, stimulates production of Mos through a series of steps, not fully known (dashed line indicates unknown intermediates). (Redrawn from ref. 31.) The right panel shows a synthetic positive feedback system²² in which the promoter region *tetreg* controls expression of the tetracycline-responsive transactivator (rtTA); rtTA activates *tetreg*, completing the positive feedback loop. **d**, Observed bistability in the synthetic positive feedback system²². Fluorescence intensities are shown for the positive feedback (autocat., autocatalytic) and unregulated (constit., constitutive) systems. The concentration of the inducer doxycycline controls the degree of positive feedback, as regulatory binding of rtTA relies on activation by the inducer. For low inducer concentrations (yellow), both the constitutive and positive feedback systems have distributions with a single peak. At higher concentrations of inducer (orange), the constitutive system remains unimodal, while the strong positive feedback causes the autocatalytic system to split into two distinct populations of cells. (Redrawn from ref. 22.)



A significant feature of positive feedback is its role in the generation of bistability, where two steady states of the system are stable (see ref. 31 for a recent review of the role of positive feedback in bistability of gene regulatory networks). The importance of positive feedback in generating multiple stable states has been analysed mathematically^{30,32} and has been implicated in the stability of the differentiated and undifferentiated states in *Xenopus* oocytes^{33,34} (Fig. 1c). Experiments with an engineered positive feedback network²² have demonstrated the existence of bistability in the system (Fig. 1d). The synthetic network was implemented in the budding yeast *Saccharomyces cerevisiae*, and consisted of a tetracycline-responsive transactivator (rtTA) that activated its own promoter. As in the negative feedback experiments,

the reporter protein GFP was fused to the transactivator to allow observation of expression levels through fluorescence microscopy. As predicted by the accompanying mathematical model²², the resulting distributions were bimodal: there were two distinct subpopulations of cells, with one group expressing small amounts of the protein, whereas the other expressed large amounts.

In the synthetic positive feedback network²², the partitioning of the cells into two subpopulations was not permanent; this was attributed to fluctuations in the network that were large enough to cause spontaneous transitions from one state to the other. Only transitions from the low-expression state to the high-expression state were observed experimentally, but there is no theoretical reason

precluding transitions in the opposite direction, and it may be chance that no such transitions occurred during the period of observation. It is well known that noise can drive a system back and forth between two stable states, and the average time it takes for such a transition to occur is called the 'escape time'. The escape time is a function of both the stability of the states and the size of the fluctuations.

In the case of *Xenopus* oocytes, differentiation is an irreversible process: once the egg cells mature, they are never observed to change back to the immature state. This indicates that the escape time of the maturation system is either infinite, or so long that it is effectively infinite; that is, over the lifetime of the organism, there is a negligible chance of making a spontaneous transition back to the immature state. (An alternative possibility in a complex network of this sort is that once maturation is achieved, the system parameters change in such a way that the bistability is irreversibly eliminated.) The fact that the synthetic positive feedback network⁴² made transitions on a significantly shorter timescale suggests that it was either subject to greater noise, or that its expression states were less stable. Controlled experiments on synthetic autoregulatory networks, combined with theoretical treatments^{16,18,35,36}, may serve to identify the precise differences between switching systems.

Toggle switch

Bistability is a minimal requirement for a network to possess memory, where the state of the network stores information about its past. When forced by a transient stimulus into one state or the other, such a system remains in that state after the transient has been removed, thus 'remembering' the stimulus event. For generating bistability, an alternative to the positive feedback network is mutual inhibition. This method of achieving bistability arises in a number of contexts: in engineering, there is the Reset–Set latch (widely known as the 'RS latch') circuit design, and switches based on mutual repression have long been suggested as a common element in gene regulatory networks¹. One such genetic switch is found in the P_R/P_{RM} region in λ phage, which acts co-repressively to control the lysis/lysogeny decision: Cro, controlled by P_R , represses P_{RM} , whereas CI, controlled by P_{RM} , represses P_R (Fig. 2a).

The principle of mutual repression was used to achieve bistability in a synthetic genetic toggle switch¹³. Its design made use of a mathematical model to deduce the parameter regimes required for bistability and robust switching. These criteria included the use of strong and balanced constitutive promoters, effective transcriptional repression, the formation of protein multimers, and similar protein degradation rates for the two main components.

An example of one of the toggle switch designs is shown in Fig. 2b. In this version of the toggle, the *lac* gene is under the control of the P_{Ls1con} promoter, whereas the *cl* gene is controlled by the P_{trc-2} promoter; the lactose repressor (LacI) protein represses P_{trc-2} , and the CI protein represses P_{Ls1con} . Experimentally, switching between the two states was induced by the transient application of either a chemical or thermal stimulus (Fig. 2c). The chemical inducer was isopropyl- β -D-thiogalactopyranoside (IPTG), which binds to LacI tetramers and renders them effectively unable to repress P_{trc-2} . A temperature-sensitive version of the CI protein was used, so that protein denaturation increased with temperature, allowing a thermal stimulus to eliminate the active CI in the system.

Logic gates

The concept of engineered gene circuits has led to a formulation based on logic gates and their associated truth tables, with resulting schematics that are the direct analogue of electronic circuit diagrams^{37,38}. One such description defines the inputs to a regulated promoter as the protein/inducer pair, and the output as 'on' if the gene downstream of the promoter is being transcribed and 'off' otherwise (Fig. 3a). For example, consider the arabinose operon, which is induced by a complex consisting of AraC dimers and the chemical arabinose. The inputs are AraC proteins and arabinose, and the arabinose

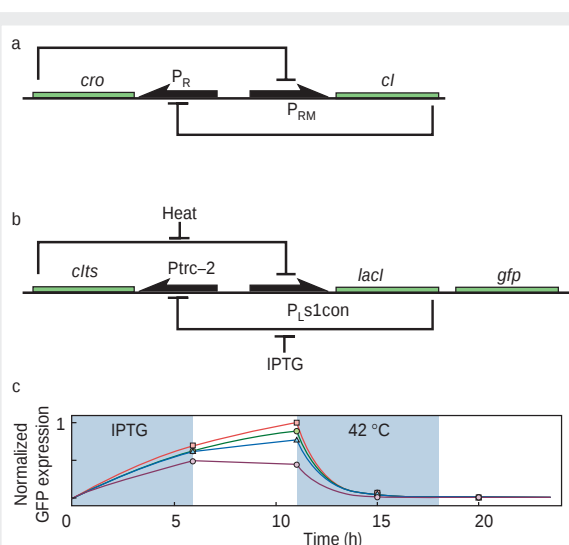


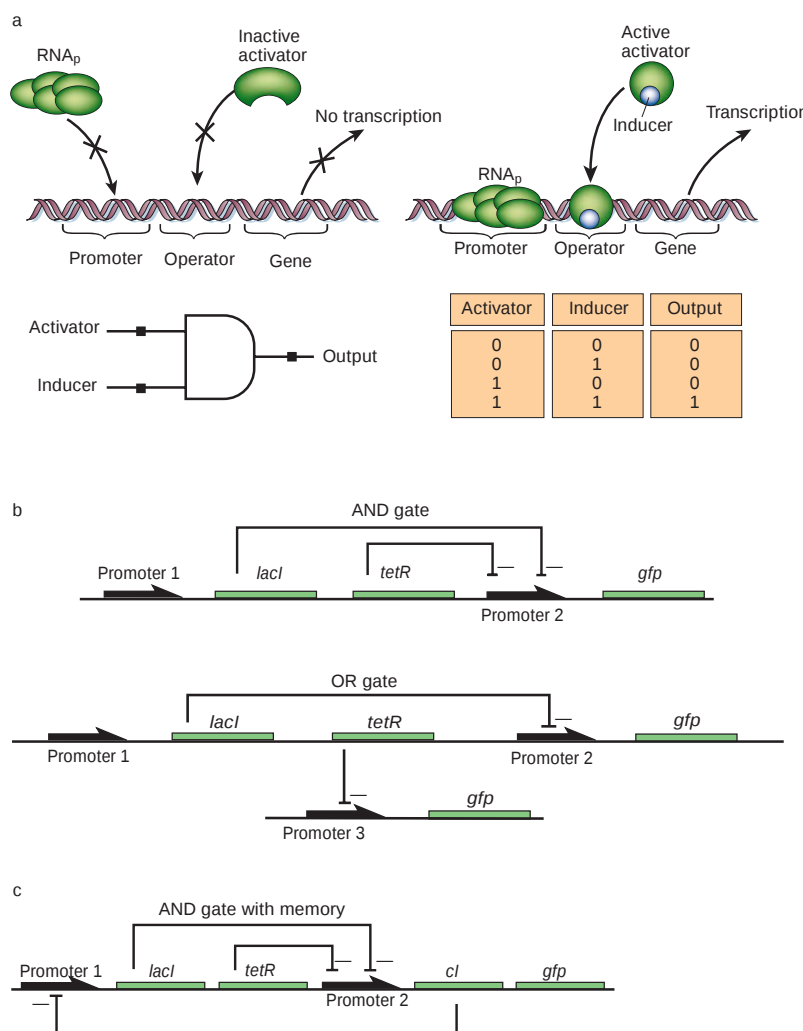
Figure 2 Natural and synthetic co-repressive switches. **a**, In this natural switch from the bacteriophage λ , the promoters P_{RM} and P_R are each repressed by the product of the other: P_{RM} controls expression of the gene *cl*, and the protein CI represses P_R , whereas P_R controls expression of the gene *cro*, and the protein Cro represses P_{RM} . **b**, The synthetic genetic toggle switch¹³ uses the promoter P_{trc-2} to control the production of a temperature-sensitive version of the CI protein (expressed by *clts*); CI acts to repress the promoter P_{Ls1con} . Conversely, P_{Ls1con} controls transcription of the gene *lacI*, whose product LacI (lactose repressor) represses P_{trc-2} . **c**, Experimental results showing bistability of a genetic toggle switch in *Escherichia coli*¹³. The response of green fluorescent protein (GFP) is shown, which corresponds to expression of the *cl* gene. Shaded regions indicate periods of induced switching. The cells were forced to the high-GFP state by exposure to isopropyl- β -D-thiogalactopyranoside (IPTG), which eliminates the repressive effect of LacI; note that the cells remain in the high state after the inducer is removed. The population was then forced into the low-GFP state by induction with a thermal pulse, which eliminated the active CI in the system by increasing the denaturation rate of the temperature-sensitive CI protein. Upon returning to the base temperature, the cells once again remain in the switched state after the induction is removed. The different coloured lines represent different plasmid strains implementing slight variations of the toggle. (Redrawn from ref. 13.)

promoter is on only if AraC and arabinose are present, and is off otherwise. In the parlance of electrical engineering, the truth table for such a system of inputs and outputs leads to a logical AND gate. Simple logic gates, such as AND gates and OR gates (see Fig. 3b), can be combined to yield circuits of any given complexity, and indeed a central focus of ref. 38 was the formulation of an engineering circuit discipline with a simulation package for analysing the resulting gene circuits³⁸.

Although the description of logic gates in terms of protein-chemical inputs is one possible approach, a complementary formulation involves defining two external chemicals as the input signals. For example, consider the schematic for an alternative AND gate depicted in Fig. 3b (F. J. Isaacs, C. R. Cantor and J.J.C., manuscript in preparation). The circuitry is such that the first promoter directs the polycistronic transcription of the *lac* and *tet* genes, and the second promoter is engineered to be repressed by either LacI or TetR. When the two chemicals IPTG and anhydrotetracycline (aTc) are present, the LacI and TetR repressors, respectively, are inactivated and the *gfp* gene downstream of promoter 2 is transcribed. Thus, the circuit forms an AND gate as both chemical inputs (IPTG and aTc) must be present for promoter 2 to be on (GFP expressed).

A central theme in gene circuit design is that the simpler 'fundamental' circuits form the basis for more complex designs. For example, memory can be added to the AND network by letting the second promoter direct the production of a third protein capable of

Figure 3 Logic gates. **a**, Genetic and electronic circuit diagrams for an AND gate using proteins and inducers for the inputs, and the state of the gene (on or off) as the output. The corresponding truth table elucidates the logic of the AND gate. **b**, An alternative AND gate uses two inducers as inputs and expression of the green fluorescent protein (*gfp*) gene as the output. The *lacI* and *tetR* genes (encoding tetracycline and lactose repressor, respectively) are expressed polycistronically by a constitutive promoter. If either LacI or TetR bind to the second promoter, the expression of the *gfp* gene is turned off. Because both of the inducers isopropyl- β -D-thiogalactopyranoside (IPTG) and anhydrotetracycline (aTc) are needed to prohibit the repression of the second promoter by LacI and TetR, the circuit forms an AND gate. The OR gate similarly uses the polycistronic expression of *lacI* and *tetR* from the first promoter, but differs in that only LacI represses the second promoter, while TetR represses a third promoter. Because the presence of either inducer (or both) leads to the expression of the *gfp* gene, the circuit forms an OR gate. **c**, Memory in the AND gate is achieved by inserting an additional *cl* gene under the control of the second promoter, and using a CI-repressible first promoter. Once the system is switched into the on state, CI represses the production of LacI and TetR, keeping the system in that state, regardless of the subsequent levels of inducers.



repressing the first promoter (Fig. 3c). This could be realized by inserting the *cl* gene (as in the toggle switch) alongside *gfp* as a polycistron, and having the first promoter be repressed by CI (for example, the P_L promoter used in the toggle switch). In this case, once the system is switched to the on state by the simultaneous presence of IPTG and aTc, it will maintain this state regardless of the subsequent concentrations of inducers applied, because the expressed CI will repress the production of LacI and TetR. In this way, the system has memory such that the presence of the on state indicates that, at some point in the past, both IPTG and aTc were present simultaneously.

Recently, a new approach involving 'combinatorial synthesis' was used to generate a myriad of logical gene circuits³⁹. This approach involved the clever use of subcloning and ligation, whereby 15 distinct promoter-gene units were constructed such that subsequent ligation of a mixture of the units yielded a library of three-gene networks. Specifically, the initial promoter-gene constructs incorporated uniquely designed *Bgl*I restriction sites in the polymerase chain reaction primers. This constrained the networks to the structure P_L -*lacI*- P_J -*lacI*- P_K -*tetR*, where P_L , P_J and P_K were each one of the five promoters P^L_1 (repressed by LacI), P^L_2 (repressed by LacI), P^T (repressed by TetR), P^A (repressed by CI) or P^A_2 (activated by CI). For measurement, a fourth transcriptional unit consisting of P^A -*gfp* was incorporated in each plasmid, so that the input-output characteristics consisted of IPTG and aTc as inputs, and GFP fluorescence as output. The plasmid library was then transformed into *Escherichia*

coli and grown under the four input conditions, with or without IPTG and with or without aTc. Analysis entailed the search for specific gene circuits in which the output fluorescence was a function of both inducers, and the result was a collection of logical circuits that included NAND, NOR and NOT IF gates.

Repressilator

Oscillations are used in engineering control systems as central 'clocks' to synchronize behaviour, and many multicellular organisms use a form of cellular 'clock' to coordinate their behaviour over the course of the day-night cycle^{40,41}. These circadian rhythms manifest themselves in the periodic variation of concentrations of particular proteins in the cell. Although the precise molecular mechanism underlying even the most basic circadian rhythm is not fully understood, a number of general models describing these important rhythms have been proposed (refs 42–47, and see review in this issue by Goldbeter, pages 238–245).

When designing synthetic networks, an alternative to building a system that reproduces exact natural mechanisms is to generate systems that exhibit similar behaviour. This approach was used¹⁴ to address the question of cellular oscillations, whereby a synthetic network (the 'repressilator') was produced that generated self-sustaining periodic oscillations in the concentrations of three proteins in a bacterial cell. The design operates on the same general principle as a ring oscillator in microelectronics. Accordingly, the repressilator

network architecture is cyclic (Fig. 4a), in which the LacI protein represses the promoter for the *tet* gene, the TetR protein represses the promoter for the *cl* gene, and the CI protein represses the promoter for the *lac* gene. As depicted in Fig. 4b, the network produced roughly sinusoidal oscillations in protein concentrations, observed by parallel expression of the reporter protein GFP.

As in the case of the toggle switch, a mathematical model was instrumental in the process of designing the repressilator. Although the ring network architecture is theoretically capable of sustaining oscillations^{14,48}, not all parameter choices give rise to oscillatory solutions. The modelling work indicated that oscillations were favoured by high protein synthesis and degradation rates, large cooperative binding effects, and efficient repression. These theoretical conclusions led to specific design choices: strong and tightly repressible hybrid promoters were selected, and the effective protein degradation rates were increased by *ssrA* tagging, whereby proteins are modified by the addition of an amino acid sequence which makes them targets for proteases in the cell.

An engineered circuit approach to sources of noise

Because the biochemical rates of transcription and translation are proportional to the number of promoter sites and messenger RNA molecules, these rates are typically small and imply relatively infrequent transcriptional and translational events compared with other interactions within the cell (for example, protein–protein interactions). In biochemistry, such infrequent events lead naturally to large fluctuations, and these fluctuations are known as internal noise because they originate from the underlying biochemical reactions rather than from some external perturbation or detection limitation.

The notion that such internal noise could be important in the choice of a developmental pathway for an organism has induced a flurry of modelling research devoted to the role of fluctuations in gene regulation (refs 25, 49–51; and see review in this issue by Arkin and co-workers, pages 231–237). Recently, theoretical models have been combined with engineered gene networks to elucidate the dominant source of internal noise in a single-gene network^{20,23}. Given the two-step process of transcription and translation, the specific goal of this work was to determine their relative contribution to the fluctuations observed in the expressed protein concentrations within a cell.

Modelling work predicted that the random variation in expression from a single gene should scale linearly with the translational rate and be independent of the transcriptional rate²⁰. Experimentally, point mutations were used to independently vary the transcriptional and translational rates, and the results were consistent with the theoretical predictions: the fluctuations in the expressed protein concentrations were observed to increase linearly with the translational efficiency while showing only a mild increase with the transcriptional efficiency²³. Of particular note was the finding that the size of the fluctuations induced in the translational step was inversely proportional to the mRNA half-life, implying that fast mRNA turnover could be a means of mitigating noise. Because fast mRNA turnover increases the cellular energy requirement for protein production, the authors speculated that the evolution of gene regulation might entail a compromise between noise reduction and energy conservation.

The importance of tightly controlled amounts of cellular protein has led other researchers to model how specific network properties might act to decrease or utilize fluctuations^{24,46,47,50,51}. One such study focused on a linear array of genes forming a network where each gene activates its nearest downstream neighbour²⁴. The central finding was that cascades can act as attenuators for a noisy input signal, thus elucidating their potential importance in cell-wide signal transduction. Other studies showed how a model circadian network can function reliably in the presence of internal noise^{46,47}. Although the underlying genetic architecture for the various known circadian systems has not been deduced, these networks seem to involve both positive- and negative-control elements⁴⁰. This information was used to construct a generic model capable of oscillations that are

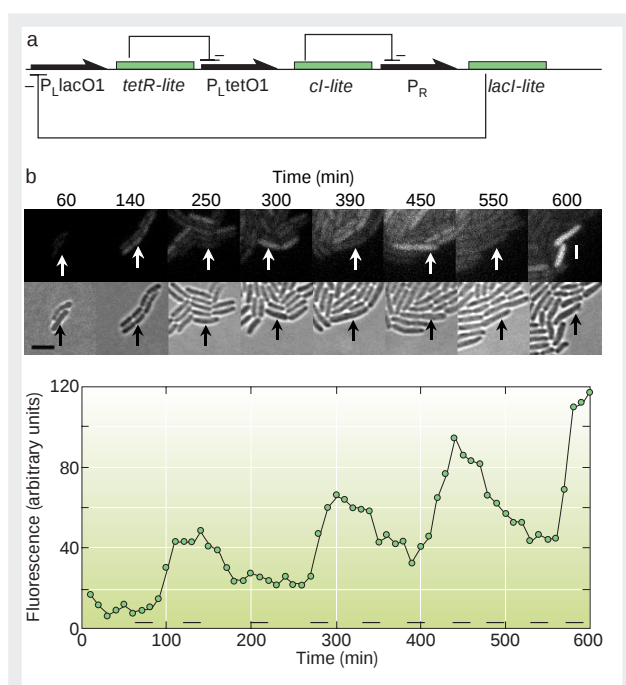


Figure 4 Synthetic transcriptional oscillator (the repressilator¹⁴). **a**, Network architecture. The synthetic system consists of three gene-promoter pairs arranged in a ring, such that each promoter's gene product represses the next promoter in the cycle. The promoter P_{lacO1} controls transcription of the gene *tetR-lite*, and the tetracycline repressor protein TetR represses the next promoter in the sequence, P_{tetO1} . P_{tetO1} controls the transcription of *cl-lite*, and the protein Cl represses the promoter P_R . Finally, P_R controls the expression of *lacI-lite*, and the lactose repressor protein LacI represses P_{lacO1} , completing the cycle. Note that the suffix 'lite' in the gene names refers to the presence of *ssrA* tags, which increase the degradation rate of the proteins. **b**, Experimental results showing oscillations in the repressilator¹⁴. The growth and timecourse of green fluorescent protein (GFP) expression was recorded for an individual *Escherichia coli* cell containing the repressilator plasmids; the cell was tracked using fluorescence (upper images) and bright-field (lower images) microscopy. Scale bar, 4 μ m. The plot below these images shows a time series of fluorescence intensities, clearly indicating oscillatory behaviour in the cell. Bars at the bottom of the plot indicate the timing of cell division events. Note that the period of the oscillations is longer than the cell division time. (Adapted from ref. 14.)

resistant to fluctuations. The study also provided evidence that circadian oscillations might actually be enhanced by noise. This leads to the conjecture that the circadian circuitry has evolved to both reduce internal fluctuations and to exploit the residual noise that cannot be fully eliminated.

The cascade and circadian network models described above provide clear theoretical predictions that can be tested systematically with engineered gene circuits. For example, cascades can be synthetically designed and the noise properties elucidated with single-cell microscopy. Similarly, the network underlying the proposed circadian oscillator could be built using an autocatalytic feedback loop as the primary network element¹⁸.

Intercell signalling system

The use of signals to coordinate the behaviour of many individual devices is crucial in microelectronics and robotics, and cells also display a significant ability to communicate, both within multicellular organisms and within populations of unicellular organisms. Because cellular membranes act to isolate the cell from its environment, such communication generally relies on specialized chemicals that either pass through the membrane (through passive diffusion or active transport) or activate membrane-spanning receptors on the exterior of the cell.

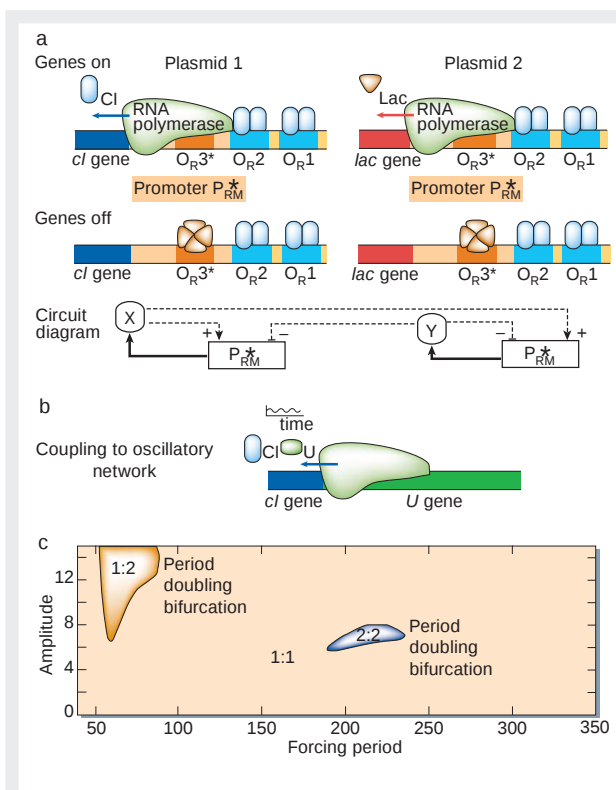


Figure 5 Synthetic oscillator design and synchronization properties. **a**, Schematic for the synthetic gene oscillator. The P_{RM}^* promoter is a mutant of the P_{RM} promoter that naturally exists in the virus λ phage⁶⁵. In its natural state, the state of the virus is regulated by CI dimers, which bind to the three right operator sites O_{R1} , O_{R2} and O_{R3} . In our design, the O_{R3} operator is replaced with an operator region O_{R3}^* , which has an affinity only for LacI (lactose repressor) tetramers. The depicted position of the Lac operator site is for illustrative purposes only, as the ideal placement of the operator may be upstream of O_{R1} and O_{R2} . **b**, The synthetic oscillator is coupled to the host genome by inserting the *cl* gene adjacent to an oscillating gene product in the host. **c**, The resonance regions are depicted in a plot of the drive amplitude versus the drive frequency. Within these regions, the period of the synthetic network oscillations is entrained to that of the external drive. (Panels **a–c** redrawn from ref. 60.)

Experiments⁵² have demonstrated the feasibility of sending signals between synthetic regulatory networks residing in different cells. This work made use of a well-studied, natural intercell signalling system, the quorum sensing pathway in the bacterium *Vibrio fischeri*^{53,54}. In quorum sensing, bacteria regulate their behaviour based on the density of bacteria present nearby. Each bacterium secretes a signalling molecule (a homoserine lactone referred to as an ‘autoinducer’), which passes through the cellular membrane in both directions. When many bacteria are present, the concentration of autoinducer reaches levels sufficient to activate a regulatory protein, LuxR, which then binds to the *lux* operator region and activates the expression of a suite of genes causing the bacterium to become luminescent.

In the synthetic system⁵², two populations of cells were engineered: ‘sender’ cells containing an autoinducer synthase (LuxI) under the control of a chemically inducible promoter; and ‘receiver’ cells containing a reporter protein (GFP) controlled by the *lux* operator region. When the sender cells were induced to express LuxI, autoinducer was produced and diffused into the extracellular environment; the autoinducer then entered the receiver cells, and stimulated production of GFP by activating the *lux* region.

A simple example of coordinated behaviour is the synchronization of oscillators, and a recent modelling study⁵⁵ considered the use of the above-described intercell signalling system to synchronize a

population of synthetic genetic relaxation oscillators¹⁸. Relaxation oscillators exhibit rapid transitions followed by periods of slow change, and previous theoretical work^{56,57} showed that such oscillators are more readily synchronized than their more smoothly varying sinusoidal counterparts, such as the repressilator. The theoretical analysis indicated that rapid synchronization could be achieved by coupling each cell’s production of autoinducer to its oscillatory phase. Experimentally verifying this prediction would be an interesting application of synthetic gene networks, as would a direct comparison of the synchronization behaviour of relaxation and sinusoidal genetic oscillators.

Applications

The above examples of engineered gene circuits serve to highlight how an integrated approach that combines computational modelling with experimental molecular biology can lead to insights into some of the basic modules that comprise complex, naturally occurring gene networks. The long-term goal of such work is to assemble increasingly complete models of the behaviour of natural systems, while maintaining at each stage the ability to test models in a tractable experimental system. An important complementary aspect of this approach is that the designer gene circuits which form the sub-modules will probably have important biotechnological applications in their own right. In this context, engineered gene networks represent a first step towards logical cellular control, whereby biological processes can be manipulated or monitored at the genetic level.

From the construction of a simple set of genetic building-block circuits (such as toggle switches and oscillators), one can imagine the design and construction of integrated biological circuits capable of performing increasingly elaborate functions. An integrated biological circuit could, like electronic control circuits, possess data-processing and storage circuitry, as well as input–output components necessary for sensing and affecting its environment. Ultimately, synthetic gene circuits encoded into DNA might be ‘downloaded’ into cells creating, in effect, a ‘wet’ nano-robot. These cellular robots could be used for a variety of functions, including *in vivo* biosensing, autonomously synthesizing complex biomaterials, executing programmed cell death, and interfacing with microelectronic circuits by transducing biochemical events to and from the electronics.

As an example of an integrated biological circuit, consider a recently engineered oncolytic adenovirus capable of selectively killing tumour cells⁵⁸. In most tumour cells, the p53 gene network does not function properly, and this dysfunction leads to an unusually low amount of the tumour-suppressing p53 protein⁵⁹. The engineered adenovirus is capable of detecting the presence or absence of p53, and executing a specific task depending on the p53 ‘state’ of the cell. If the amount of p53 is normal, a viral promoter controlling the inhibition of replication is turned on and viral replication is halted. But if the amount of p53 is low, the virus detects the abnormal cell and replication proceeds along with the expression of viral proteins that lead to cell lysis and the spread of the adenovirus to other potentially cancerous cells.

Examples of other complex network-control schemes are provided by several recent modelling studies that focus on the utility of coupling designer gene networks to native cellular processes. One such study explored the coupling of an oscillating synthetic network to intrinsic cell-generated oscillations⁶⁰ (Fig. 5). This work provided design strategies for entraining and amplifying oscillations in cellular protein concentrations. Such control could prove useful in the design of networks that interact with cellular processes that require precise timing. Along these lines, seminal developments in the modelling of the cell-division cycle^{61–63} could be coupled to the oscillator model, allowing for the design of protein delivery schemes that are signalled by the cellular growth cycle.

Another modelling study explored the utilization of engineered gene networks in the reverse engineering of large-scale gene

regulatory networks⁶⁴. Here, the central idea is that small, engineered networks can be inserted into cells to provide a controlled perturbation mechanism for ongoing gene-expression experiments. One then tracks how the perturbation affects the genes in a naturally occurring network, and this information can be used to deduce the network topology. This method may prove useful in identifying and validating specific drug targets and in unravelling the effects of chemical compounds.

By reducing the complexity of the systems under study, synthetic gene networks offer the ability to gain a detailed understanding of the mechanisms involved in gene regulation. As our grasp of the fundamental principles of gene regulation improves, we will be able to design and study increasingly complex systems. The conjunction of the advanced experimental techniques of molecular biology with the mathematical tools of nonlinear dynamics and statistical physics provides an exciting opportunity for rapid advances in the understanding and control of cellular behaviour. □

doi:10.1038/nature01257

1. Monod, J. & Jacob, F. General conclusions: telenomic mechanisms in cellular metabolism, growth, and differentiation. *Cold Spring Harb. Symp. Quant. Biol.* **26**, 389–401 (1961).
2. Glass, L. & Kauffman, S. A. The logical analysis of continuous, non-linear biochemical control networks. *J. Theor. Biol.* **39**, 103–129 (1973).
3. Savageau, M. A. Comparison of classical and autogenous systems of regulation in inducible operons. *Nature* **252**, 546–549 (1974).
4. Kauffman, S. A. The large-scale structure and dynamics of gene control circuits: an ensemble approach. *J. Theor. Biol.* **44**, 167–190 (1974).
5. Glass, L. Classification of biological networks by their qualitative dynamics. *J. Theor. Biol.* **54**, 85–107 (1975).
6. Glass, L. Combinatorial and topological methods in nonlinear chemical kinetics. *J. Chem. Phys.* **63**, 1325–1335 (1975).
7. Savageau, M. A. *Biochemical System Analysis* (Addison Wesley, Reading, 1976).
8. Goodwin, B. C. *Analytical Physiology of Cells and Developing Organisms* (Academic, London, 1976).
9. Tyson, J. J. & Othmer, H. G. The dynamics of feedback control circuits in biochemical pathways. *Prog. Theor. Biol.* **5**, 1–62 (1978).
10. McAdams, H. H. & Shapiro, L. Circuit simulation of genetic networks. *Science* **269**, 650–656 (1995).
11. McAdams, H. H. & Arkin, A. Towards a circuit engineering discipline. *Curr. Biol.* **10**, R318–R320 (2000).
12. Reintz, J. & Vaisnys, J. R. Theoretical and experimental analysis of the phage lambda genetic switch implies missing levels of co-operativity. *J. Theor. Biol.* **145**, 295–318 (1990).
13. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403**, 339–342 (2000).
14. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–338 (2000).
15. Becskei, A. & Serrano, L. Engineering stability in gene networks by autoregulation. *Nature* **405**, 590–593 (2000).
16. Hasty, J., Pradines, J., Dolnik, M. & Collins, J. J. Noise-based switches and amplifiers for gene expression. *Proc. Natl Acad. Sci. USA* **97**, 2075–2080 (2000).
17. Smolen, P., Baxter, D. A. & Byrne, J. H. Mathematical modeling of gene networks. *Neuron* **26**, 567–580 (2000).
18. Hasty, J., Isaacs, F., Dolnik, M., McMillen, D. & Collins, J. J. Designer gene networks: towards fundamental cellular control. *Chaos* **11**, 207–220 (2001).
19. Hasty, J., McMillen, D., Isaacs, F. & Collins, J. J. Computational studies of gene regulatory networks: in *numero* molecular biology. *Nature Rev. Genet.* **2**, 268–279 (2001).
20. Thattai, M. & van Oudenaarden, A. Intrinsic noise in gene regulatory networks. *Proc. Natl Acad. Sci. USA* **98**, 8614–8619 (2001).
21. Simpson, M. L., Saylor, G. S., Fleming, J. T. & Applegate, B. Whole-cell biocomputing. *Trends Biotechnol.* **19**, 317–323 (2001).
22. Becskei, A., Séraphin, B. & Serrano, L. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *EMBO J.* **20**, 2528–2535 (2001).
23. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nature Genet.* **31**, 69–73 (2002).
24. Thattai, M. & van Oudenaarden, A. Attenuation of noise in ultrasensitive signaling cascades. *Biophys. J.* **82**, 2943–2950 (2002).
25. McAdams, H. H. & Arkin, A. Stochastic mechanisms in gene expression. *Proc. Natl Acad. Sci. USA* **94**, 814–819 (1997).
26. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C51 (1999).
27. Lauffenburger, D. A. Cell signaling pathways as control modules: complexity for simplicity? *Proc. Natl Acad. Sci. USA* **97**, 5031–5033 (2000).
28. Palsson, B. O. & Lightfoot, E. N. Mathematical modeling of dynamics and control in metabolic networks. *J. Theor. Biol.* **113**, 279–298 (1985).

29. Thomas, R. The role of feedback circuits: positive feedback circuits are a necessary condition for positive real eigenvalues of the Jacobian matrix. *Ber. Bessenges. Phys. Chem.* **98**, 1148–1151 (1994).
30. Thomas, R., Thieffry, D. & Kaufman, M. Dynamical behaviour of biological regulatory networks-I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bull. Math. Biol.* **57**, 247–276 (1995).
31. Ferrell, J. E. Jr Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Curr. Opin. Cell Biol.* **14**, 140–148 (2002).
32. Keller, A. Model genetic circuits encoding autoregulatory transcription factors. *J. Theor. Biol.* **172**, 169–185 (1995).
33. Ferrell, J. E. Jr Xenopus oocyte maturation: new lessons from a good egg. *BioEssays* **21**, 833–842 (1999).
34. Thomas, R., Thieffry, D. & Kaufman, M. Dynamical behaviour of biological regulatory networks-II. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. *Bull. Math. Biol.* **57**, 247–276 (1995).
35. Aurell, E., Brown, S., Johanson, J. & Sneppen, K. Stability puzzles in phase λ . *Phys. Rev. E* **65**, 051914-1–051914-9 (2002).
36. Bialek, W. in *Advances in Neural Information Processing* Vol. 13 (eds Leen, T. K., Dietterich, T. G. & Tresp, V.) 103–109 (MIT Press, Cambridge, MA, 2001).
37. Weiss, R. & Basu, S. The device physics of cellular logic gates. First Workshop on Non-Silicon Computing <http://www-2.cs.cmu.edu/~phoenix/nsc1/paper/3-2.pdf> (2002).
38. Weiss, R. *Cellular Computation and Communications Using Engineered Genetic Regulatory Networks*. Thesis, Massachusetts Institute of Technology (2001).
39. Guet, C., Elowitz, M., Hsing, W. & Leibler, S. Combinatorial synthesis of genetic networks. *Science* **296**, 1466–1470 (2002).
40. Dunlap, J. Molecular bases for circadian clocks. *Cell* **96**, 271–290 (1999).
41. Panda, S., Hogenesch, J. B. & Kay, S. A. Circadian rhythms from flies to human. *Nature* **417**, 329–335 (2002).
42. Leloup, J. C. & Goldbeter, A. A model for circadian rhythms in *Drosophila* incorporating the formation of a complex between the PER and TIM proteins. *J. Biol. Rhythms* **13**, 70–87 (1998).
43. Tyson, J. J., Hong, C. I., Thron, C. D. & Novak, B. A simple model of circadian rhythms based on dimerization and proteolysis of PER and TIM. *Biophys. J.* **77**, 2411–2417 (1999).
44. Leloup, J. C. & Goldbeter, A., Modeling the molecular regulatory mechanism of circadian rhythms in *Drosophila*. *BioEssays* **22**, 84–93 (2000).
45. Roussel, M. R., Gonze, D. & Goldbeter, A. Modeling the differential fitness of cyanobacterial strains whose circadian oscillators have different free-running periods: comparing the mutual inhibition and substrate depletion hypotheses. *J. Theor. Biol.* **205**, 321–340 (2000).
46. Barkai, N. & Leibler, S. Biological rhythms: circadian clocks limited by noise. *Nature* **403**, 267–268 (2000).
47. Vilar, J. M. G., Kueh, H. Y., Barkai, N. & Leibler, S. Mechanisms of noise-resistance in genetic oscillators. *Proc. Natl Acad. Sci. USA* **99**, 5988–5992 (2002).
48. Smith, H. Oscillations and multiple steady states in a cyclic gene model with repression. *J. Math. Biol.* **25**, 169–190 (1987).
49. Arkin, A., Ross, J. & McAdams, H. H. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ -infected *Escherichia coli* cells. *Genetics* **149**, 1633–1648 (1998).
50. Kepler, T. & Elston, T. Stochasticity in transcriptional regulation: origins, consequences and mathematical representations. *Biophys. J.* **81**, 3116–3136 (2001).
51. Paulsson, J., Berg, O. G. & Ehrenberg, M. Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. *Proc. Natl Acad. Sci. USA* **97**, 7148–7153 (2000).
52. Weiss, R. & Knight, T. F. in *DNA6: Sixth International Meeting on DNA Based Computers* (Leiden, The Netherlands, 2000).
53. Fuqua, C., Winans, S. & Greenberg, E. P. Census and consensus in bacterial ecosystems: the LuxR-LuxI family of transcriptional regulators. *Annu. Rev. Microbiol.* **50**, 727–751 (1996).
54. Bassler, B. L. How bacteria talk to each other: regulation of gene expression by quorum sensing. *Curr. Opin. Microbiol.* **2**, 582–587 (1999).
55. McMillen, D., Kopell, N., Hasty, J. & Collins, J. J. Synchronizing genetic relaxation oscillators by intercell signaling. *Proc. Natl Acad. Sci. USA* **99**, 679–684 (2002).
56. Somers, D. & Kopell, N. Rapid synchronization through fast threshold modulation. *Biol. Cybern.* **68**, 393–407 (1993).
57. Wang, D. L. in *Proc. 15th Annu. Conf. Cognit. Sci. Soc.* 1058–1063 (Lawrence Erlbaum Assoc., Hillsdale, NJ, 1993).
58. Ramachandra, M. et al. Re-engineering adenovirus regulatory pathways to enhance oncolytic specificity and efficacy. *Nature Biotech.* **19**, 1035–1041 (2001).
59. Vogelstein, B., Lane, D. & Levine, A. J. Surfing the p53 network. *Nature* **408**, 307–310 (2000).
60. Hasty, J., Dolnik, M., Rottschäfer, V. & Collins, J. J. A synthetic gene network for entraining and amplifying cellular oscillations. *Phys. Rev. Lett.* **88**, 148101-1–148101-4 (2002).
61. Novak, B. & Tyson, J. J. Modeling the control of DNA replication in fission yeast. *Proc. Natl Acad. Sci. USA* **94**, 9147–9152 (1997).
62. Svecizer, A., Csikasz-Nagy, A., Györfy, B., Tyson, J. J. & Novak, B. Modeling the fission yeast cell cycle: quantized cycle times in *wee1* Δ mutant cells. *Proc. Natl Acad. Sci. USA* **97**, 7865–7870 (2000).
63. Chen, K. C., Csikasz-Nagy, A., Györfy, B., Novak, B. & Tyson, J. J. Kinetic analysis of a molecular model of the budding yeast cell cycle. *Mol. Biol. Cell* **11**, 369–391 (2000).
64. Tegner, J., Yeung, M. K. S., Hasty, J. & Collins, J. J. Reverse engineering gene networks—integrating genetic perturbations with dynamical modeling. *Proc. Natl Acad. Sci. USA* (submitted).
65. Ptashne, M. et al. How the λ repressor and cro work. *Cell* **19**, 1–11 (1980).

Acknowledgements

This work was supported by the Defense Advanced Research Projects Agency (DARPA), Office of Naval Research (ONR), National Science Foundation (NSF) BioQuBIC, the Fetzer Institute, and Natural Sciences and Engineering Research Council of Canada (NSERC).

Control, exploitation and tolerance of intracellular noise

Christopher V. Rao*, Denise M. Wolf† & Adam P. Arkin*†‡

Departments of Bioengineering* and Chemistry†, University of California, and ‡Physical Biosciences Division, Lawrence Berkeley National Laboratory, Howard Hughes Medical Institute, 1 Cyclotron Road, MS 3-144, Berkeley, California 94720, USA (e-mail: c_rao@lbl.gov; dmwolf@lbl.gov; aparkin@lbl.gov)

Noise has many roles in biological function, including generation of errors in DNA replication leading to mutation and evolution, noise-driven divergence of cell fates, noise-induced amplification of signals, and maintenance of the quantitative individuality of cells. Yet there is order to the behaviour and development of cells. They operate within strict parameters and in many cases this behaviour seems robust, implying that noise is largely filtered by the system. How can we explain the use, rejection and sensitivity to noise that is found in biological systems? An exploration of the sources and consequences of noise calls for the use of stochastic models.

“For it is simply a fact of observation that the guiding principle in every cell is embodied in a single atomic association existing only in one copy (or sometimes two) — and a fact of observation that it results in producing events which are paragons of orderliness [...] the situation is unprecedented, it is unknown anywhere else except in living matter.”

Erwin Schrödinger *What is Life?*
(Cambridge University Press, 1944)

When Erwin Schrödinger wrote *What is Life?*, he was interested in whether new physical laws were necessary to describe biological systems. He was acutely concerned with how “a single group of atoms existing in one copy produces orderly events”. Although the molecular basis of genetics did not require new physical laws, how cells function and process information when the underlying molecular events are random still remains an open question. Gene expression, for example, involves a series of single-molecule events and belies a deterministic description. As each of these molecular events is subject to significant thermal fluctuations, gene expression is best viewed as a stochastic process. Even in cases where population measurements are regular and reproducible, single-cell measurements often display significant heterogeneity¹. Overall, these observations suggest that the molecular events underlying cellular physiology are subject to fluctuations and have led to the proposal of a stochastic model^{2–4} for gene expression and biochemistry in general. Other cellular processes influenced by noise include ion-channel gating⁵, neural firing⁶, cytoskeleton dynamics⁷ and motors⁸, although here we focus primarily on the role of noise in intracellular networks.

How do we explain the complex, highly orchestrated and robust physiology of the cell when the underlying molecular events are basically random? Despite the stochastic function of the foundations of regulatory circuits within cells, most cellular events are ordered and precisely regulated. Development in *Caenorhabditis elegans* is so regular that we can trace the differentiated state of nearly every cell⁹. One example where the transition from disorder to order has been measured is in *Drosophila melanogaster* embryos¹⁰. Although the anterior-to-posterior gradient of the maternal morphogen Bicoid in

D. melanogaster embryos displays significant variability, the profile of the *hunchback* gap gene, regulated by Bicoid, is precise. The need for order has led to the proposal that robustness is an intrinsic property of intracellular networks^{11,12}.

Although most cellular processes are ordered, not all noise is rejected. Cell fate and population heterogeneity is viewed increasingly as a noise-driven process. In the phage lambda infection process, which is governed by the lysis–lysogeny decision circuit, only a fraction of infecting phage chooses to lyse the cell. The remainder become dormant lysogens awaiting bacterial stress signals to enter the production phase of their life cycle¹³. Another example of population heterogeneity can be found in the soil-growing bacterium *Bacillus subtilis*, which responds to environmental stress with an arsenal of probabilistically invoked survival strategies. *B. subtilis* can become motile and swim towards new food sources, secrete degradative enzymes to scavenge resources, secrete antibiotics to eliminate competitors, produce stress-resistant spores, or become competent for genetic transformation¹⁴. The particular fate of each cell seems random, although biased by environmental and intercellular signals. Still more examples of population heterogeneity include differentiation of progenitor haematopoietic stem cells¹⁵, non-genetic individuality in bacterial chemotaxis¹⁶, and epigenetic inheritance and incomplete penetrance of transgenes in mice¹⁷. However, even heterogeneity is ordered; once a particular fate is chosen, the resulting process is tightly controlled.

Does the noise manifested as random cell fate and population heterogeneity help or hurt the organism, or does it have an indifferent effect? In at least some cases, randomness and heterogeneity seem to be a boon to survival. Phase variation in pathogenic bacteria, where cells alternate randomly between expressing certain genes and silencing others, is thought to be a form of cultivated noise¹⁸. Type 1 pili expression in uropathic *Escherichia coli*^{18–21}, pili expression in *Neisseria gonorrhoeae*²², polysaccharide intercellular adhesion synthesis in *Staphylococcus epidermidis*²³, lipopolysaccharide epitope expression in *Haemophilus influenzae*²⁴, and capsular polysaccharide expression in *Vibrio vulnificus*²⁵ are just a few examples of this common mode of control²⁶. Even though the molecular events leading to phase variation seem random in the individual, regulatory factors tune the variation to ensure mean levels of

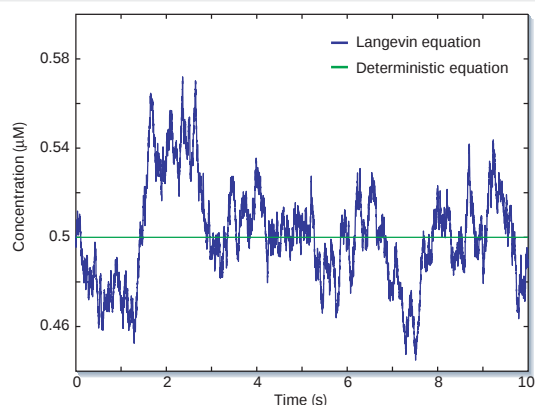


Figure 1 Comparison of the deterministic and stochastic solution for an isomerization reaction with dissociation constant $K_d = 1$. The deterministic solution predicts a constant equilibrium (green line). The stochastic solution obtained by solving a Langevin equation includes fluctuations about the equilibrium concentration. The Langevin equation was solved using a first-order Euler method.

heterogeneity for the population. Environmental factors can shape population diversity, presumably allowing for an adaptive response to the conflicting demands of offence (infection of the host) and defence (immune system recognition and destruction of the pathogen)^{13,26}.

Although examples of tightly ordered or potentially noise-exploiting cellular processes abound, how cells are able to reign in biochemical noise remains unknown. Where does noise arise in the cell? By what means do regulatory networks attenuate this noise? And how and why do networks exploit noise? These questions present one of the most challenging and fascinating problems for systems (if not all) biologists, as they open questions in physiology, development and evolutionary biology. The answer likely resides in the complex networks that underlie cellular physiology. Computational models are the ideal tool for such investigations, because they allow us to express formally the current state of knowledge about network composition and structure, and to explore network dynamics. These tools allow us to test and generate hypotheses about the fundamental operating principles of a network and the sources and consequences of intracellular noise, something not possible with qualitative arguments.

Modelling tools

Biochemical reactions are described traditionally in terms of kinetic rates that describe how the concentrations of the various species (for example, proteins or metabolites) in a cell (or test tube) change with time. The reaction rates are embodied by rate laws such as mass action or Michaelis–Menten kinetics, and the biochemical dynamics are described with differential equations. A typical form of the equation is

$$\frac{dC(t)}{dt} = vr(C)$$

where the variables $C(t)$, t , v and $r(C)$ represent the concentrations, time, stoichiometric matrix and the rate law, respectively. Implicit in the above formulation is the assumption that the cell is well mixed and homogenous. This assumption is not limiting as the model can be formulated with a spatial component that describes phenomena such as cytoplasmic heterogeneity, compartmentalization, diffusion and wave phenomena. Literally hundreds of software packages (both commercial and freeware) are available to construct and solve, either analytically or numerically, equations of these forms^{27,28}.

These models are deterministic; if the starting conditions are fixed, then the future evolution is also fixed precisely. Despite this, it is

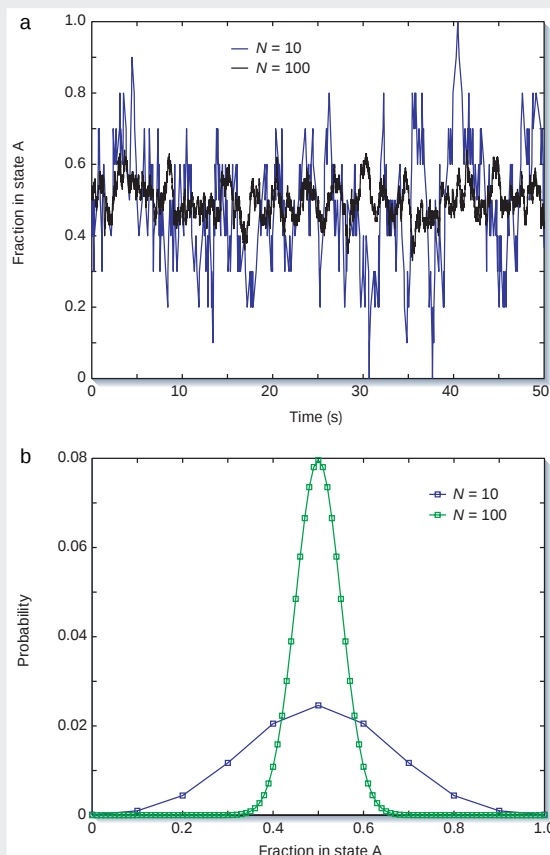


Figure 2 A comparison of the isomerization reaction with 10 and 100 molecules using a discrete stochastic model with $k_1 = 1 \text{ s}^{-1}$ and $k_2 = 1 \text{ s}^{-1}$. **a**, The sizes of fluctuations decrease as the number of molecules increases. Simulations were performed using the Gillespie algorithm. **b**, The steady-state probability density function. As the number of molecules increases, the density becomes sharper. The figure shows a plot of the analytic solution for the steady-state master equation. The distribution is given by the expression

$$p(j) = \binom{n}{j} \frac{k_1^j k_2^{n-j}}{(k_1 + k_2)^n}$$

This discussion is adapted from ref. 62.

possible to study the effects of noise to a first approximation using bifurcation and spectral analysis. These approaches assume noise arises from an exogenous source and tacitly ignore intrinsic fluctuations in pathway (for example, a noisy ligand signal is assumed and fluctuations arising in the signal-transduction cascade are ignored).

Molecular fluctuations can be incorporated explicitly by including random variables (or rather stochastic processes) in the model. The easiest approach is to append a noise term to the end of the differential equation

$$\frac{dC(t)}{dt} = vr(C) + x(t)$$

where $x(t)$ is the additive (white) noise term. The equation above is often referred to as the Langevin equation or a stochastic differential equation²⁹. The appeal of the Langevin approach is that it builds on the deterministic formulation (Fig. 1).

While many algorithms exist for simulating the Langevin equation³⁰, often one calculates the probability density function instead.

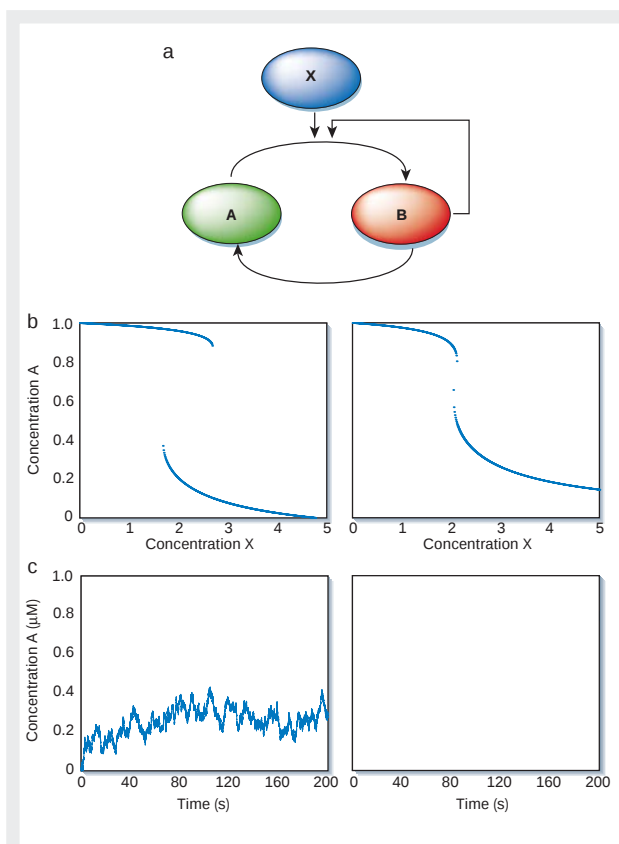


Figure 3 Switches and chattering. **a**, A simple reaction network with positive feedback produces a switch. **b**, The steady-state behaviours of two switches as a function of the signal X . The left curve is a hysteresis whereas the right curve is ultrasensitive. Differences between the two curves result from the use of different kinetic parameters in the model. **c**, The dynamic behaviour of the two switches subject to similar noisy X signals was simulated using the Gillespie algorithm. The hysteresis of the first switch provides a buffer so that the switch is robust to noise. The second switch, which is ultra-sensitive and lacks such a buffer, is sensitive to noise and subject to accidental switching (indicated by arrow). Hysteretic switches provide one mechanism to reduce switching chatter.

The Fokker–Planck (or Kolmogorov’s forward equation) describes the evolution of the probability density function

$$\frac{\partial p(C,t)}{\partial t} = -\nabla \cdot [vr(C)p(C,t)] + \frac{1}{2} \sum_{i,j} \frac{\partial^2 \sigma_{ij} p(C,t)}{\partial C_i \partial C_j}$$

where $p(C,t)$ is the probability density function and the matrix σ_{ij} is the covariance of the noise process $x(t)$. The quantity $p(C,t)\partial C$ is the probability of finding a cell with a concentration of a certain chemical between C and $C + \partial C$ at time t . One advantage of working with the Fokker–Planck equation is that it is possible to analyse the model. Tools such as sensitivity analysis and bifurcation theory are applicable. However, for systems involving more than a few species, it is impossible to solve the Fokker–Planck equation, even numerically. Most researchers analyse these models using Monte–Carlo methods, where one solves the Langevin equation many times and then uses statistics to estimate the probability density function. Compared to deterministic equations, Monte–Carlo methods are time consuming when simulating many molecules and reactions, although they currently are the only option for complex (that is, realistic) models.

Implicit, however, in either differential or Langevin equations is a continuous description of molecular species, where the dynamics are cast in terms of infinitesimal changes in concentration. This

description is limiting when modelling processes involve a few molecules, discrete structures or single-reaction events such as the binding of a transcription factor to its cognate promoter.

Recent research in biological noise has been directed towards modelling molecular species (such as proteins, messenger RNA and ribosomes) as discrete entities using elements of probability theory. In this framework, reaction events replace reaction rates, and each distinct reaction event is explicitly modelled. The likelihood of a reaction event (for example, a protein undergoing a transition) is analogous to a reaction rate. Rather than referring to a differential rate, we assign a probability that the protein will undergo a transition in an infinitesimal amount of time. Consider a protein existing in two states A or B. We can now write a differential equation of the form

$$\frac{dP(n_a, n_b; t)}{dt} = -(k_1 n_a + k_2 n_b)P(n_a, n_b; t) + k_1(n_a + 1)P(n_a + 1, n_b - 1; t) + k_2(n_b + 1)P(n_a - 1, n_b + 1; t)$$

which describes how the probability $P(n_a, n_b; t)$ that n_a proteins exist in state A and n_b proteins exist in state B changes as a function of time. The parameters k_1 and k_2 denote the likelihood of an A-to-B and B-to-A transition, respectively. This equation is called a master equation and describes what statisticians call a birth–death process. It also defines a homogenous Markov chain, and is actually no different mathematically than the equations used commonly in sequence analysis, population biology and theoretical genetics. The master equation is linear and, from a mathematical perspective, it is about as simple an equation as one can hope for. The caveat is that the equation is large, so large you never want to write it down. Because the problem above is simple, we can calculate an analytic solution for the steady-state probability distribution (Fig. 2).

As with the Fokker–Planck equation, the master equation is deterministic in the sense that if the starting probabilities are fixed, then the future probabilities are fixed. The main difference between the two formulations is how the species are represented: the description is continuous in the Fokker–Planck equation, but discrete in the master equation. When modelling only a few reacting molecules, the discrete representation is believed to be more accurate than the continuous representation. However, as the number of molecules increases this difference becomes less significant. In fact, the master equation is asymptotically equivalent to the Langevin equation

$$\frac{dC(t)}{dt} = vr(C) + v\sqrt{r(C)}x(t)$$

where $x(t)$ is a unit white-noise process^{31–33}. This equation predicts that the relative magnitude of the molecular fluctuations scales roughly as the inverse square root of the number of reacting molecules.

Rarely do we work directly with the master equation, as many equations are necessary to model systems involving more than ten reactions or species. For example, the master equation requires ten thousand equations to describe a three-step linear pathway involving one hundred molecules, as an equation is necessary to account for each possible combination of molecules. Rather than enumerate every state (the tumour suppressor p53 has at least 11 phosphorylation and acetylation sites, implying 2^{11} distinct states for the monomer and potentially 2^{44} states for the tetramer³⁴), it is easier to simulate the random evolution of the system and use Monte–Carlo approaches. This solution was formulated by Gillespie³⁵, who proposed a simple, elegant algorithm for simulating stochastic kinetics. This task is then repeated many times to estimate the relevant probabilities and statistics. Although this procedure may be time consuming, it is far easier than forming and then solving the master equation. In Gillespie’s algorithm, the time for the next reaction event is calculated and the system is updated accordingly in an iterative manner.

An alternative approach to the Gillespie algorithm for stochastic simulation is the StochSim algorithm^{36,37}. In StochSim, the master equation is discretized to facilitate numerical approximations of the transition probabilities describing the evolution of the biochemical dynamics. The approach is reminiscent of an explicit forward Euler method for solving differential equations. It is not 'exact'; the error is proportional to the size of the time increments. If, however, we choose small time increments, the error is negligible and StochSim is asymptotically equivalent to the Gillespie algorithm. An alternate approach to the master equation, tailored for diffusion processes in complex geometries such as ion transport in synapses, is MCELL (<http://www.mcell.cnsl.salk.edu/>). MCELL uses a ray-tracing algorithm for tracking molecular motion and interactions. For systems in thermodynamic equilibrium, the problem can often be recast in terms of the Boltzmann equation, and Monte-Carlo solutions can be obtained using the Metropolis algorithm and variants thereof.

There are still many unresolved issues regarding stochastic simulation, computational efficiency being the most pressing. Although a few strategies have been proposed to increase the efficiency of the Gillespie algorithm^{38,39}, there are currently no satisfactory approaches for simulating processes concurrently across multiple scales of time, space or concentration. An alternative approach is to separate timescales explicitly and reduce the model by singular perturbations⁴⁰. Yet other approach is to construct hybrid models involving continuous and discrete representations⁴¹. Both these approaches require direct intervention by the modeller — a cumbersome and sometimes impossible task. The long-term goal is to develop algorithms that do this both automatically and adaptively. But the challenge to multiscale simulation is rare events. How do we simulate the rare events of interest without wasting computational resources simulating frequent events that are irrelevant to the question being asked? One can envision algorithms analogous to adaptive methods used to solve stiff differential equations, whose realization will likely involve a time discretization similar to the StochSim algorithm.

How is a modeller to choose between modelling approaches — an implicit or explicit treatment of noise, a continuous or discrete representation of molecules? When simulating processes that involve only a few molecules, discrete stochastic models are superior to continuous models. However, in many processes, there are many copies of some species and few of others. In these circumstances, it is not always clear which approach is better. Detailed mechanisms are easier to include using discrete representations. For example, we can explicitly model the structure of the chromosome, transcription, translation, ribosome and polymerase queues on mRNA and DNA respectively, and events such as convergent transcription⁴². The disadvantage of discrete models is that they are more difficult to formulate, test and solve computationally. As multiscale approaches for simulating stochastic processes are desperately lacking, personal proclivities currently dictate the choice of approach, as modelling and simulation are, at this stage, more art than science.

Noise analysis

The modelling tools described above allow us to address questions concerning intracellular noise. A few of these questions include where noise arises in cells, how pathways function robustly in spite of noise, how molecular noise can selectively generate population heterogeneity, and how cells potentially exploit noise.

Origin of noise

To study the origins of noise in gene expression, McAdams and Arkin⁴ proposed a stochastic model for gene expression in prokaryotes. Their model suggests that proteins are produced in random bursts. As a single mRNA transcript can produce multiple copies of a protein, protein translation amplifies transcriptional noise. Numerous other models have further validated and extended this hypothesis by analysing the mechanisms contributing to noise in gene expression^{43–45}. As an experimental verification, van Oude-

naarden and colleagues studied how the frequencies of transcription and translation contribute to variability in gene expression by measuring expression of a green fluorescent protein (GFP) marker⁴⁶. Their results provided explicit evidence that most noise arises during translation.

What fraction of noise is attributable to fluctuations in gene expression and what fraction to external (or extrinsic) fluctuations arising from other cellular components? To discriminate between the two sources, Elowitz and colleagues⁴⁷ measured differential expression of distinguishable cyan and yellow fluorescent protein markers under the control of identical promoters. The degree of correlation in a single cell provides a measure of discrimination; as fluctuations in gene expression increase, the degree of correlation decreases. By varying levels of gene expression using a lac promoter, they showed that fluctuations in gene expression decrease as the expression increases. Likewise, extrinsic noise decreases with increased levels of gene expression, although remarkably it first passes through a maximum at intermediate levels of expression. In other words, at low levels of expression both forms of noise are present, whereas extrinsic noise dominates at intermediate levels, and both forms are absent at high levels. It was also shown that noise has a genetic component; *recA* mutants are twice as noisy as their wild-type counterparts.

Noise control mechanisms

Many researchers have found it useful to invoke analogies from signal processing when investigating noise^{48,49}. From this perspective, a pathway is viewed as an analog filter and is classified in terms of its frequency response. Cascades and relays such as two-component systems and the mitogen-activated protein kinase pathway have inherent noise-rejecting properties⁵⁰. In terms of signal processing, these pathways function as low-pass filters, as they transduce low-frequency signals whereas high-frequency signals are attenuated. In fact, most physical systems attenuate high-frequency noise on input signals because of inherent time lags and delays. But noise also arises in the pathway as a result of internal molecular fluctuations, and we cannot simply ignore this noise or separate noise in the signal from that in the pathway. Where this type of separation has been attempted, it has been observed that in certain network topologies, such as cascades, there seems to be a trade-off between noise attenuation of an input signal and inherent noise generated at each step of the pathway. van Oudenaarden and colleagues examined this trade-off for cascade structures, and suggested that there is an optimal cascade length for attenuating noise⁵¹. This analysis illustrates how conclusions regarding noise may be derived from deterministic models through indirect analysis.

Perhaps the simplest and most common noise-attenuating regulatory mechanism is negative feedback. The principle of negative feedback is to measure whether the behaviour is acceptable, and to make corrections based on the 'error' between the desired and measured behaviours. In the fields of engineering and economics, it is well known that negative feedback is necessary to operate robustly in an uncertain environment. Not surprisingly, feedback is ubiquitous in biology as it provides a simple mechanism to attenuate the effects of noise^{52–54}. In terms of its signal-processing capabilities, a simple negative feedback loop functions as a low-pass filter. Becksei and colleagues⁵⁵ demonstrated this effect by constructing a negative feedback module in *E. coli*. Their experiments showed that constitutive expression of the GFP is highly variable, in terms of the measured fluorescence intensity, whereas the addition of the negative feedback using the tetracycline repressor significantly reduces the measured variability, as expected.

Whereas simple negative feedback results in a low-pass filter, another type of feedback — integral feedback — shapes a band-pass filter. Integral feedback is a form of negative feedback that uses an internal memory state to amplify intermediate frequencies and attenuate low and high frequencies. Bacterial chemotaxis is an example of a system using integral feedback⁵⁶. Here, integral feedback

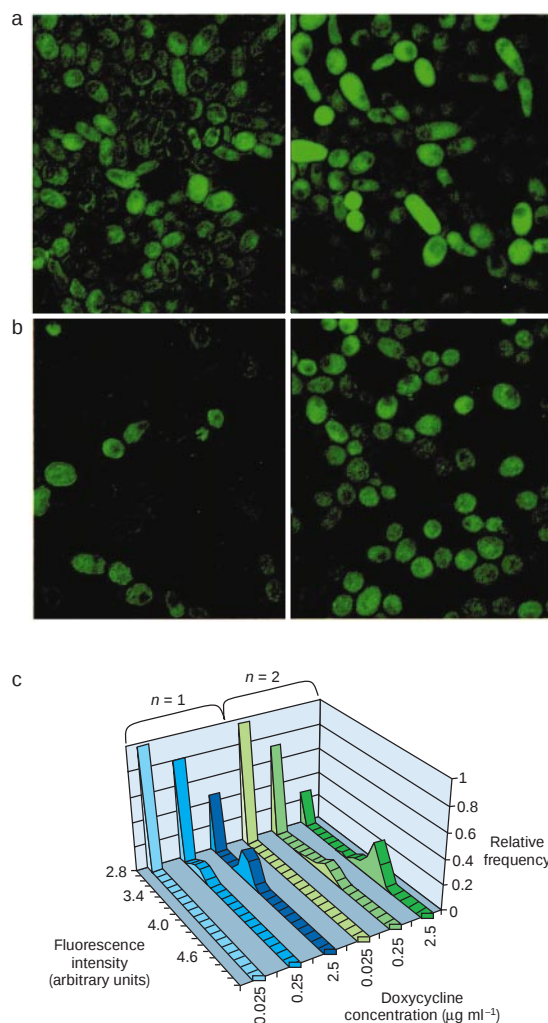


Figure 4 Construction of the synthetic positive feedback loop of Becksei and colleagues⁶⁶. **a**, Constitutive expression of the tetracycline-responsive transactivator (rtTA). The degree of activation, measured by expression of the green fluorescent protein (GFP), is proportional to the amount of inducer, doxycycline, added. Cells were induced with 0.25 $\mu\text{g ml}^{-1}$ (left) and 2.5 $\mu\text{g ml}^{-1}$ (right) doxycycline. **b**, rtTA under the control of a positive feedback loop. In this set up, rtTA regulates the expression of GFP and itself. Cells were induced with 0.25 $\mu\text{g ml}^{-1}$ (left) and 2.5 $\mu\text{g ml}^{-1}$ (right) doxycycline. At low concentrations of inducer, the response is heterogeneous. **c**, Distribution of fluorescence level in the positive feedback loop construction as a function of inducer concentration and gene copy number (n). The heterogeneous response is indicated by the bimodal distribution. (Images courtesy of A. Becksei and *EMBO Journal*⁶⁶).

measures temporal changes in chemical concentrations, rather than steady-state changes, and results in biased motion towards attractants and robust adaptation.

In addition to intrinsic chemical damping, negative feedback and integral feedback, many other simple mechanisms attenuate noise in systems. One example is redundancy mechanisms such as gene dosage and parallel cascades^{57,58}. These mechanisms attenuate the effects of noise by increasing the likelihood of gene expression or establishing a consensus from multiple signals. Another example is regulatory checkpoints⁵⁹. Best characterized in the cell cycle and flagellar biosynthesis, checkpoints ensure that each step in a pathway is successfully completed before proceeding with the next step. Yet

another example is kinetic proofreading in protein translation, where mechanisms are in place to correct possible errors⁶⁰.

Noise amplification and exploitation

Complementary work has focused attention on cellular processes that amplify or exploit noise in some sense, rather than just controlling or eliminating it. These processes fall into two classes — mechanisms that give rise to population heterogeneity (and thus diversity), and mechanisms that use noise to attenuate noise. Isogenic heterogeneity seems to arise from a noisy step in the commitment portion of an otherwise ordered process. One example is the genetic circuit governing development in phage lambda, where it was proposed⁴² that molecular fluctuations cause an initially homogenous population to partition into a heterogeneous lytic and lysogenic population. The basic mechanism governing the decision circuit involves two antagonistic feedback loops — crossed repressive feedback loops generate a switch, and molecular fluctuations partition the population statistically so that individuals may (by chance) follow one path or the other. These results illustrate how intrinsic molecular noise is used to generate diversity.

The *fim* network regulates phase variation of type 1 pili in uropathic *E. coli*. Type 1 pili, which are adhesive organelles expressed on the surface of the cell, are virulence factors in urinary tract infections^{18,20}. A mechanism was proposed⁶¹ where the system components (invertible DNA element and the global regulators and invertases that act stochastically upon it) realize a number of devices that together transduce environmental signals into inversion probabilities and thus the heterogeneity level of the population, presumably creating piliated populations in the bladder and unpiliated populations outside the host. This network includes a switch based on the ratio of regulatory proteins, a temperature tuning device capable of reading the temperature and increasing piliation at mammalian body temperature, and a delay line using feedback as memory to prevent rapid cycling between ON and OFF switching states (discussed below). This system provides an example of how integrated regulatory modules in a network can function to both shape and filter noise, thereby creating environmentally tuned heterogeneity in a cell population.

The *fim* network seems to include a delay that decreases the sensitivity of the switch to noise. Switches may be sensitive to both noise and ‘chatter’ (Fig. 3). Chattering arises commonly in engineering, where noisy signals may cause switches to rapidly turn off and on, and it was proposed that the flagellar motor in bacterial chemotaxis possesses a mechanism to prevent chatter⁶². The expected fluctuations in the response regulator CheY were shown not correlate with the switching behaviour, suggesting that the flagellar motor has a mechanism that decreases sensitivity to noise in CheY. Latter experiments showed that the flagellar switch may possess a hysteresis⁶³, one mechanism known to reduce chatter.

Feedback can also amplify the effects of noise by autocatalytic mechanisms^{64,65} (that is, positive feedback). In an experimental study by Becksei and colleagues⁶⁶, a synthetic positive feedback loop in yeast was constructed using the tetracycline transactivator and a GFP marker (Fig. 4). In this system, activation of the feedback loop is variable and randomness at the single-cell level leads to a mixed colony of cells.

In addition to generating heterogeneous populations, cells also use noise to filter noise. Whereas in most systems noise degrades a signal, noise actually enhances a signal when certain nonlinear effects are present. One example is stochastic resonance⁶⁷; numerous examples of this exist in biology, such as electroreceptors in paddlefish⁶⁸, mechanoreceptors in the tail fins of crayfish⁶⁹ and hair cells in crickets⁷⁰. It has also been suggested that noise can potentially increase sensitivity in certain signalling cascades⁷¹.

Complex interactions and multiple feedback loops

Some of the elementary mechanisms for noise attenuation, amplification and exploitation enumerated above present the illusion of

tractability (that is, they appear simple and readily identifiable). However, elementary mechanisms typically do not function in isolation, but rather interact in complex networks involving multiple feedback loops. These regulatory networks can produce diverse phenomena ranging from switches to memory to oscillators^{72,73}. Although it is straightforward to understand how a single feedback loop shapes noise, it is far more difficult to understand the composite behaviour of multiple mechanisms interconnected in complex architectures.

It is for these interactions that computational models are most useful. For example, the network that controls circadian rhythms consists of multiple, complex, interlocking feedback loops. Many researchers have investigated the mechanisms for noise resistance in circadian rhythms, using both deterministic and stochastic models^{74–77}. General models of chemical oscillators are sensitive to kinetic parameters. However, the proposed mechanisms for circadian rhythms produce regular oscillations in the presence of noise. Remarkably, the stochastic model is able to produce regular oscillations when the deterministic models do not⁷⁶, suggesting that the regulatory networks may utilize molecular fluctuations to their advantage.

Other examples of complex networks functioning in the presence of noise are early expression of *hox* genes⁷⁸ and bacterial chemotaxis^{79,80}. As with the previous example, noise attenuation arises from the systematic properties of the network rather than from a single mechanism. What specific mechanisms confer robust functionality in the presence of noise? Apparently, noise attenuation arises from complex mechanisms involving multiple feedback loops. Although theoretical and computational tools exist for analysing the properties of a given network, no good theory exists (and perhaps never will) for identifying all possible mechanisms that generate robust networks.

It is clear that large, complex networks are able to function reliably despite inherent noise attributable to molecular fluctuations. Although simple, specific mechanisms to explain this phenomenon can be elusive, robustness has been hypothesized as an intrinsic property of intracellular networks. In two landmark papers, Leibler and colleagues showed that the chemotaxis pathway in *E. coli* is robust^{11,81}; the pathway is functional for a wide range of enzymatic activities and protein concentrations. Other examples of robustness include developmental processes^{12,82} and phage lambda regulation⁸³. Although robustness is often studied independently of noise, the two problems are not distinct. When studying robustness, the typical question is how sensitive the behaviour of a network is to the parameters in the model. As these parameters are subject to fluctuations, a noise-resistant network is likely to be robust. But a network that is insensitive to the kinetic parameters may still be sensitive to molecular noise, as internal and external noise are rarely parameterized explicitly in these models. A comprehensive investigation of robustness needs to account explicitly for noise.

Beneath the noise

The studies described above highlight the need for understanding the role of noise in biology. We are still far from answering the question “How does order arise from disorder?” but we are beginning to get a glimpse of some of the mechanisms by which cells control and exploit noise.

Considerations of noise and robustness offer insight into the design and function of intracellular networks^{84–86}. In particular, what design features or constraints are necessary for pathways to function robustly in the presence of noise? Design and function often imply teleological arguments. Rather than teleology, the hypothesis is that the function of a network and the need for robustness impose constraints on its design and canalize evolution. For example, protein function often implies a specified chemistry, such as a membrane protein having a hydrophobic region, a soluble globular protein having a hydrophobic core, and the active site of serine protease containing a Ser-His-Asp catalytic triad. We expect similar

constraints on networks; in particular, function imposes a specific regulatory and information structure. We do not suggest that networks are designed optimally for fitness, but rather that certain design features are necessary for a stable phenotype. These design constraints allude to a theoretical biology distinct from physics and chemistry, more akin to engineering than the new physical laws that Schrödinger originally envisioned. □

doi:10.1038/nature01258

- Ko, M. S., Nakauchi, H. & Takahashi, N. The dose dependence of glucocorticoid-inducible gene expression results from changes in the number of transcriptionally active templates. *EMBO J.* **9**, 2835–2842 (1990).
- Berg, O. G. A model for the statistical fluctuations of protein numbers in a microbial population. *J. Theor. Biol.* **71**, 587–603 (1978).
- Ko, M. S. A stochastic model for gene induction. *J. Theor. Biol.* **153**, 181–194 (1991).
- McAdams, H. H. & Arkin, A. Stochastic mechanisms in gene expression. *Proc. Natl Acad. Sci. USA* **94**, 814–819 (1997).
- White, J. A., Rubinstein, J. T. & Kay, A. R. Channel noise in neurons. *Trends Neurosci.* **23**, 131–137 (2000).
- Allen, C. & Stevens, C. F. An evaluation of causes for unreliability of synaptic transmission. *Proc. Natl Acad. Sci. USA* **91**, 10380–10383 (1994).
- van Oudenaarden, A. & Theriot, J. A. Cooperative symmetry-breaking by actin polymerization in a model for cell motility. *Nature Cell Biol.* **1**, 493–499 (1999).
- Simon, S. M., Peskin, C. S. & Oster, G. F. What drives the translocation of proteins? *Proc. Natl Acad. Sci. USA* **89**, 3770–3774 (1992).
- Sternberg, P. W. & Felix, M. A. Evolution of cell lineage. *Curr. Opin. Genet. Dev.* **7**, 543–550 (1997).
- Houchmandzadeh, B., Wieschaus, E. & Leibler, S. Establishment of developmental precision and proportions in the early *Drosophila* embryo. *Nature* **415**, 798–802 (2002).
- Barkai, N. & Leibler, S. Robustness in simple biochemical networks. *Nature* **387**, 913–917 (1997).
- von Dassow, G., Meir, E., Munro, E. M. & Odell, G. M. The segment polarity network is a robust developmental module. *Nature* **406**, 188–192 (2000).
- Prashne, M. A. *Genetic Switch: Phage Lambda and Higher Organisms* (Cell Press, Blackwell Scientific Publications, Cambridge, MA, 1998).
- Msadek, T. When the going gets tough: survival strategies and environmental signaling networks in *Bacillus subtilis*. *Trends Microbiol.* **7**, 201–207 (1999).
- Mayani, H., Dragowska, W. & Lansdorp, P. M. Lineage commitment in human hemopoiesis involves asymmetric cell division of multipotent progenitors and does not appear to be influenced by cytokines. *J. Cell. Physiol.* **157**, 579–586 (1993).
- Spudis, J. L. & Koshland, D. E. Jr Non-genetic individuality: chance in the single cell. *Nature* **262**, 467–471 (1976).
- Morgan, H. D., Sutherland, H. G., Martin, D. I. & Whitelaw, E. Epigenetic inheritance at the agouti locus in the mouse. *Nature Genet.* **23**, 314–318 (1999).
- Connell, I. et al. Type 1 fimbrial expression enhances *Escherichia coli* virulence for the urinary tract. *Proc. Natl Acad. Sci. USA* **93**, 9827–9832 (1996).
- Abraham, J. M., Freitag, C. S., Clements, J. R. & Eisenstein, B. I. An invertible element of DNA controls phase variation of type 1 fimbriae of *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **82**, 5724–5747 (1985).
- Mulvey, M. A., Schilling, J. D., Martinez, J. J. & Hultgren, S. J. Bad bugs and beleaguered bladders: interplay between uropathogenic *Escherichia coli* and innate host defenses. *Proc. Natl Acad. Sci. USA* **97**, 8829–8835 (2000).
- Sauer, F. G., Mulvey, M. A., Schilling, J. D., Martinez, J. J. & Hultgren, S. J. Bacterial pili: molecular mechanisms of pathogenesis. *Curr. Opin. Microbiol.* **3**, 65–72 (2000).
- Mehr, I. J. & Seifert, H. S. Differential roles of homologous recombination pathways in *Neisseria gonorrhoeae* pilin antigenic variation, DNA transformation and DNA repair. *Mol. Microbiol.* **30**, 697–710 (1998).
- Ziebuhr, W. et al. A novel mechanism of phase variation of virulence in *Staphylococcus epidermidis*: evidence for control of the polysaccharide intercellular adhesion synthesis by alternating insertion and excision of the insertion sequence element IS256. *Mol. Microbiol.* **32**, 345–356 (1999).
- Peak, I. R., Jennings, M. P., Hood, D. W., Bisercic, M. & Moxon, E. R. Tetrameric repeat units associated with virulence factor phase variation in *Haemophilus* also occur in *Neisseria* spp. and *Moraxella catarrhalis*. *FEMS Microbiol. Lett.* **137**, 109–114 (1996).
- Wright, A. C., Powell, J. L., Kaper, J. B. & Morris, J. G. Jr Identification of a group 1-like capsular polysaccharide operon for *Vibrio vulnificus*. *Infect. Immun.* **69**, 6893–6901 (2001).
- Hallet, B. Playing Dr Jekyll and Mr Hyde: combined mechanisms of phase variation in bacteria. *Curr. Opin. Microbiol.* **4**, 570–581 (2001).
- Arkin, A. P. Synthetic cell biology. *Curr. Opin. Biotechnol.* **12**, 638–644 (2001).
- Slepchenko, B. M., Schaff, J. C., Carson, J. H. & Loew, L. M. Computational cell biology: spatiotemporal simulation of cellular events. *Annu. Rev. Biophys. Biomol. Struct.* **31**, 423–441 (2002).
- Gardiner, C. W. *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences* (Springer, Berlin, 1990).
- Kloeden, P. E. & Platen, E. *Numerical Solution of Stochastic Differential Equations* (Springer, Berlin, 1992).
- Gillespie, D. T. The chemical Langevin equation. *J. Chem. Phys.* **113**, 297–306 (2000).
- Gillespie, D. T. The chemical Langevin equation and Fokker-Planck equation for the reversible isomerization reaction. *J. Phys. Chem. A* **106**, 5063–5071 (2002).
- Kurtz, T. G. *Approximation of Population Processes* (SIAM, Philadelphia, 1981).
- Kohn, K. W. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell* **10**, 2703–2734 (1999).
- Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977).
- Le Novère, N. & Shimizu, T. S. STOCHSIM: modelling of stochastic biomolecular processes. *Bioinformatics* **17**, 575–576 (2001).
- Shimizu, T. S. & Bray, D. In *Foundations of Systems Biology* (ed. Kitano, H.) 213–232 (MIT Press, Cambridge, MA, 2001).

38. Gillespie, D. T. Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.* **115**, 1716–1733 (2001).
39. Gibson, M. A. & Bruck, J. Exact stochastic simulation of chemical systems with many species and many channels. *J. Phys. Chem. A* **105**, 1876–1889 (2000).
40. Rao, C. V. & Arkin, A. Stochastic chemical kinetics and the quasi steady-state assumption: application to the Gillespie algorithm. *J. Chem. Phys.* (in the press).
41. Haseltine, E. L. & Rawlings, J. B. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *J. Chem. Phys.* **117**, 6958–6969 (2002).
42. Arkin, A., Ross, J. & McAdams, H. H. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics* **149**, 1633–1648 (1998).
43. Thattai, M. & van Oudenaarden, A. Intrinsic noise in gene regulatory networks. *Proc. Natl Acad. Sci. USA* **98**, 8614–8619 (2001).
44. Kierzek, A. M., Zaim, J. & Zielenkiewicz, P. The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression. *J. Biol. Chem.* **276**, 8165–8172 (2001).
45. Kepler, T. B. & Elston, T. C. Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophys. J.* **81**, 3116–3136 (2001).
46. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & van Oudenaarden, A. Regulation of noise in the expression of a single gene. *Nature Genet.* **31**, 69–73 (2002).
47. Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. Stochastic gene expression in a single cell. *Science* **297**, 1183–1186 (2002).
48. Arkin, A. P. in *Self-organized Biological Dynamics and Nonlinear Control* (ed. Walleczek, J.) 112–144 (Cambridge Univ. Press, London, 2000).
49. Samoilov, M., Arkin, A. & Ross, J. Signal processing by simple chemical systems. *J. Phys. Chem. A* (in the press).
50. Detwiler, P. B., Ramanathan, S., Sengupta, A. & Shraiman, B. I. Engineering aspects of enzymatic signal transduction: photoreceptors in the retina. *Biophys. J.* **79**, 2801–2817 (2000).
51. Thattai, M. & Van Oudenaarden, A. Attenuation of noise in ultrasensitive signaling cascades. *Biophys. J.* **82**, 2943–2950 (2002).
52. Smolen, P., Baxter, D. A. & Byrne, J. H. Modeling transcriptional control in gene networks—methods, recent results, and future directions. *Bull. Math. Biol.* **62**, 247–292 (2000).
53. Fell, D. *Understanding the Control of Metabolism* (Portland, London, 1997).
54. Heinrich, R. & Schuster, S. *The Regulation of Cellular Systems* (Portland, London, 1996).
55. Becskei, A. & Serrano, L. Engineering stability in gene networks by autoregulation. *Nature* **405**, 590–593 (2000).
56. Yi, T. M., Huang, Y., Simon, M. I. & Doyle, J. Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proc. Natl Acad. Sci. USA* **97**, 4649–4653 (2000).
57. McAdams, H. H. & Arkin, A. It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet.* **15**, 65–69 (1999).
58. Cook, D. L., Gerber, A. N. & Tapscott, S. J. Modeling stochastic gene expression: implications for haploinsufficiency. *Proc. Natl Acad. Sci. USA* **95**, 15641–15646 (1998).
59. Hartwell, L. H. & Weinert, T. A. Checkpoints: controls that ensure the order of cell cycle events. *Science* **246**, 629–634 (1989).
60. Rodnina, M. V. & Wintermeyer, W. Ribosome fidelity: tRNA discrimination, proofreading and induced fit. *Trends Biochem. Sci.* **26**, 124–130 (2001).
61. Wolf, D. M. & Arkin, A. P. Fifteen minutes of fim: control of type 1 pili expression in *E. coli*. *Omic* **6**, 91–114 (2002).
62. Morton-Firth, C. J. & Bray, D. Predicting temporal fluctuations in an intracellular signalling pathway. *J. Theor. Biol.* **192**, 117–128 (1998).
63. Bren, A. & Eisenbach, M. Changing the direction of flagellar rotation in bacteria by modulating the ratio between the rotational states of the switch protein FliM. *J. Mol. Biol.* **312**, 699–709 (2001).
64. Ferrell, J. E. Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Curr. Opin. Cell Biol.* **14**, 140–148 (2002).
65. Hasty, J., Pradines, J., Dolnik, M. & Collins, J. J. Noise-based switches and amplifiers for gene expression. *Proc. Natl Acad. Sci. USA* **97**, 2075–2080 (2000).
66. Becskei, A., Seraphin, B. & Serrano, L. Positive feedback in eukaryotic gene networks: cell differentiation by graded to binary response conversion. *EMBO J.* **20**, 2528–2535 (2001).
67. Gammaioni, L., Hanggi, P., Jung, P. & Marchesoni, F. Stochastic resonance. *Rev. Mod. Phys.* **70**, 223–287 (1998).
68. Russell, D. F., Wilkens, L. A. & Moss, F. Use of behavioural stochastic resonance by paddle fish for feeding. *Nature* **402**, 291–294 (1999).
69. Douglass, J. K., Wilkens, L., Pantazelou, E. & Moss, F. Noise enhancement of information transfer in crayfish mechanoreceptors by stochastic resonance. *Nature* **365**, 337–340 (1993).
70. Levin, J. E. & Miller, J. P. Broadband neural encoding in the cricket cercal sensory system enhanced by stochastic resonance. *Nature* **380**, 165–168 (1996).
71. Paulsson, J., Berg, O. G. & Ehrenberg, M. Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. *Proc. Natl Acad. Sci. USA* **97**, 7148–7153 (2000).
72. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–338 (2000).
73. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403**, 339–342 (2000).
74. Barkai, N. & Leibler, S. Circadian clocks limited by noise. *Nature* **403**, 267–268 (2000).
75. Gonze, D., Halloy, J. & Goldbeter, A. Robustness of circadian rhythms with respect to molecular noise. *Proc. Natl Acad. Sci. USA* **99**, 673–678 (2002).
76. Vilar, J. M., Kueh, H. Y., Barkai, N. & Leibler, S. Mechanisms of noise-resistance in genetic oscillators. *Proc. Natl Acad. Sci. USA* **99**, 5988–5992 (2002).
77. Smolen, P., Baxter, D. A. & Byrne, J. H. Modeling circadian oscillations with interlocking positive and negative feedback loops. *J. Neurosci.* **21**, 6644–6656 (2001).
78. Kastner, J., Solomon, J. & Fraser, S. Modeling a *hox* gene network *in silico* using a stochastic simulation algorithm. *Dev. Biol.* **246**, 122–131 (2002).
79. Levin, M. D., Morton-Firth, C. J., Abouhamad, W. N., Bourret, R. B. & Bray, D. Origins of individual swimming behavior in bacteria. *Biophys. J.* **74**, 175–181 (1998).
80. Morton-Firth, C. J., Shimizu, T. S. & Bray, D. A free-energy-based stochastic simulation of the Tar receptor complex. *J. Mol. Biol.* **286**, 1059–1074 (1999).
81. Alon, U., Surette, M. G., Barkai, N. & Leibler, S. Robustness in bacterial chemotaxis. *Nature* **397**, 168–171 (1999).
82. Meir, E., von Dassow, G., Munro, E. & Odell, G. M. Robustness, flexibility, and the role of lateral inhibition in the neurogenic network. *Curr. Biol.* **12**, 778–786 (2002).
83. Little, J. W., Shepley, D. P. & Wört, D. W. Robustness of a gene regulatory circuit. *EMBO J.* **18**, 4299–4307 (1999).
84. Csete, M. E. & Doyle, J. C. Reverse engineering of biological complexity. *Science* **295**, 1664–1669 (2002).
85. Morohashi, M. *et al.* Robustness as a measure of plausibility in models of biochemical networks. *J. Theor. Biol.* **216**, 19–30 (2002).
86. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).

Computational approaches to cellular rhythms

Albert Goldbeter

Unité de Chronobiologie théorique, Faculté des Sciences, Université Libre de Bruxelles, Campus Plaine, CP 231, B-1050 Brussels, Belgium

Oscillations arise in genetic and metabolic networks as a result of various modes of cellular regulation. In view of the large number of variables involved and of the complexity of feedback processes that generate oscillations, mathematical models and numerical simulations are needed to fully grasp the molecular mechanisms and functions of biological rhythms. Models are also necessary to comprehend the transition from simple to complex oscillatory behaviour and to delineate the conditions under which they arise. Examples ranging from calcium oscillations to pulsatile intercellular communication and circadian rhythms illustrate how computational biology contributes to clarify the molecular and dynamical bases of cellular rhythms.

Rhythmic phenomena represent one of the most striking manifestations of dynamic behaviour in biological systems. In 1936, Fessard¹ published a book on the *Rhythmic Properties of Living Matter*. This book was devoted solely to the oscillatory properties of nerve cells, but it is now clear that rhythms are encountered at all levels of biological organization, with periods ranging from a fraction of a second to years^{2,3}. These rhythms find their roots in the many regulatory mechanisms that control the dynamics of living systems. Thus, at the cellular level, neural and cardiac rhythms are associated with the regulation of voltage-dependent ion channels, metabolic oscillations originate from the regulation of enzyme activity, pulsatile intercellular signals and intracellular calcium oscillations involve the control of receptor activity or transport processes, while regulation of gene expression underlies circadian rhythms.

Understanding the molecular and cellular mechanisms responsible for oscillations is crucial for unravelling the dynamics of life. When based firmly on experiments, computational biology provides an essential tool for studying these mechanisms which, because of their complexity, cannot be comprehended by sheer intuition alone. The purpose of this article is to present an overview of how models and computer simulations are used to address the origin, properties and functions of some of the main cellular rhythms.

Theoretical models for biological rhythms were first used in ecology to study the oscillations resulting from interactions between populations of predators and prey⁴. Neural rhythms represent another field where such models were used at an early stage: the formalism developed by Hodgkin and Huxley⁵ still forms the core of most models for oscillations of the membrane potential in nerve and cardiac cells⁶⁻⁸. Of more recent vintage are models for oscillations of non-electrical nature that arise at the cellular level from regulation of enzyme, receptor or gene activity (see ref. 3 for a detailed list of references).

The computational biology of these rhythms forms the core of this review. I shall consider, in turn, oscillations of intracellular calcium, pulsatile signalling in intercellular communication, and circadian rhythms. Additionally, I shall describe how computational biology can help in understanding the transition from simple periodic

behaviour to complex oscillations including bursting and chaos.

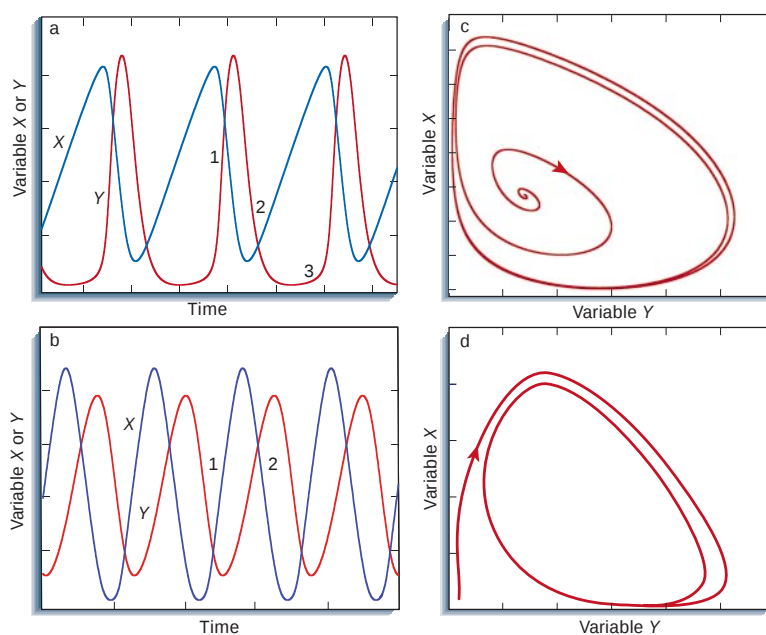
Basic phenomenology of oscillatory phenomena

In the course of time, open systems that exchange matter and energy with their environment generally reach a stable steady state. However, as shown by Glansdorff and Prigogine, once the system operates sufficiently far from equilibrium and when its kinetics acquire a nonlinear nature, the steady state may become unstable⁹. Feedback processes and cooperativity are two main sources of nonlinearity that favour the occurrence of instabilities in biological systems. When the steady state becomes unstable, the system moves away from it, often bursting into sustained oscillations around the unstable steady state (Fig. 1a, b).

In the phase space defined by the system's variables (for example, the concentrations of the biochemical species that are involved in the oscillatory mechanism), sustained oscillations correspond to the evolution towards a closed curve — the limit cycle. These oscillations are resistant to perturbations, because the limit cycle will be regained regardless of initial conditions, starting from the vicinity of the unstable state (Fig. 1c) or from outside the asymptotic, closed trajectory (Fig. 1d). Limit-cycle oscillations thus represent an example of non-equilibrium self-organization and can therefore be viewed as temporal dissipative structures⁹. The oscillations are characterized by their amplitude and by their period. A bifurcation diagram can be constructed by plotting the amplitude of the oscillations of a given variable and the steady state (stable or unstable) as a function of a control parameter (see Box 1).

Evolution towards a limit cycle is not the only possible behaviour when a steady state becomes unstable in a spatially homogeneous system. The system may evolve towards another stable steady state (when such a state exists). The most common case of multiple steady states, referred to as bistability, is of two stable steady states separated by an unstable one. This phenomenon is thought to be important in differentiation¹⁰, and was shown recently to have a role in early *Xenopus* development¹¹. When spatial inhomogeneities develop, instabilities may lead to the emergence of spatial or spatiotemporal dissipative structures⁹. These can take the form of propagating concentration waves, which are closely related to oscillations.

Figure 1 Sustained oscillations can occur in models based on positive or negative feedback. **a**, Typical oscillations obtained in models based on positive feedback. The particular oscillations shown are obtained in a two-variable model for the product-activated phosphofructokinase reaction responsible for glycolytic oscillations^{3,61}. Y represents the reaction product, and X represents the substrate of the enzyme. Similar oscillations are obtained in models for Ca^{2+} oscillations based on Ca^{2+} -induced Ca^{2+} release^{3,17} or cAMP oscillations in *Dictyostelium* amoebae^{3,28}. In the case of Ca^{2+} oscillations, Y denotes cytosolic Ca^{2+} , whereas X represents the Ca^{2+} content of intracellular stores. For cAMP oscillations in *Dictyostelium*, which rely on a mixture of positive and negative feedback (see text), X represents the fraction of active (non-desensitized) cAMP receptor, and Y represents the level of extracellular cAMP. **b**, Oscillations obtained in a five-variable model based on negative feedback for the circadian rhythmic variation of the PER protein (Y) and its mRNA (X) in *Drosophila*^{3,50}. **c**, Limit cycle in the phase plane (X , Y), corresponding to the oscillations shown in **a**. Initial conditions are such that the limit cycle here is reached from a point located in the vicinity of the unstable steady state. **d**, Limit cycle corresponding to the oscillations shown in **b**. Initial conditions are such that the limit cycle here is reached from outside. The arrows on the phase plane trajectory in **c** and **d** indicate the direction of movement along the limit cycle. Over a period, oscillations in **a** and **b** can be broken down into phases 1–3 and 1–2, respectively (see text and below). As in Fig. 3, actual scales for variables X and Y and governing kinetic equations are given in the original references indicated in the figure legends. Curves were obtained by numerical integration of these equations, using the Berkeley Madonna software. In both **a** and **b**, the rise in Y , brought about by the rise in X , leads to a drop in X . This drop is followed by a decrease in Y that eventually allows for the next rise in X . Thus, although the regulatory interactions are of opposite nature, the phase relationship between the variables is similar. One feature, however, differs between the two situations. In the case of positive feedback (**a**), the time lag associated with phase 3 results in the pulsatile nature of the oscillations: sharp peaks, or spikes, are generated at regular



intervals. Phase 3 corresponds to the pacemaker potential that brings electrically excitable cells up to the depolarization threshold beyond which an action potential is generated. This pulsatility is generally not seen in models based on negative feedback (**b**), which lack phase 3. However, negative feedback can also produce spikes at regular intervals when the oscillatory mechanism involves the passage through thresholds (such a situation producing oscillations resembling those of **a** is encountered in a model of a phosphorylation cascade for the mitotic oscillator in amphibian embryonic cells^{3,72}). In both **a** and **b**, the driving force behind sustained oscillations is found in the phase of increase of variable X , which does not require any positive feedback. In the different examples considered here, it suffices that the rise in X is brought about by some constitutive process such as substrate replenishment, receptor resensitization, refilling of Ca^{2+} stores or gene transcription.

Understanding the molecular mechanism of oscillations requires clarifying the chain of events that cause each variable of the system to periodically rise and fall. Elucidation of the underlying mechanism largely reduces to identifying the feedback processes that lie at the core of the oscillations. The latter may originate from positive (Fig. 1a) or negative (Fig. 1b) feedback, or from a mixture of both. The interplay between a large number of variables coupled through multiple regulatory interactions makes it difficult, if not impossible, to fully grasp the dynamics of oscillatory behaviour without resorting to modelling and computer simulations.

In addressing the molecular mechanism of a biological rhythm, the typical programme of computational biology consists of the following steps. First, the key variables of the phenomenon are identified, together with the nature of their interactions that form the relevant feedback loops. Second, differential equations describing the time evolution of the system are constructed. In spatially homogeneous conditions, these take the form of ordinary differential equations, whereas in the presence of diffusion, partial differential equations are used to describe the system's spatiotemporal evolution. Third, the steady state(s) admitted by these equations are determined analytically or by numerical integration.

The fourth step probes the stability properties of the steady state(s). This is generally done by using linear stability analysis. The principle of this analysis³ is to determine the evolution of infinitesimal perturbations away from the steady state: the steady state is stable when such perturbations decay in time, and unstable otherwise. When parameter values correspond to an unstable steady state,

numerical integration of the evolution equations should confirm that in the course of time the system leaves the steady state to evolve either to another, stable steady state or to sustained limit-cycle oscillations.

Using this approach, the fifth step is to determine the domains of occurrence of sustained oscillations in parameter space. Numerical solution of the kinetic equations then allows the construction of bifurcation diagrams that show how the period and amplitude vary as a function of the various parameters. Bifurcation diagrams may also be generated by means of programs (such as AUTO, developed by Doedel¹²) which are based on continuation methods. Finally, the theoretical predictions of the model, obtained by numerical simulations based on available parameter values, or else on values taken in a physiological range, are compared with experimental observations. This programme possesses its own dynamics: when the model predictions do not agree with experiments, or when new behaviours are discovered, the model must be modified accordingly.

Calcium oscillations

The three best-known examples of biochemical oscillations were found during the decade 1965 to 1975^{3,13,14}. These include the peroxidase reaction, glycolytic oscillations in yeast and muscle, and the pulsatile release of cyclic AMP (cAMP) signals in *Dictyostelium* amoebae (see below). Another decade passed before the development of Ca^{2+} fluorescent probes led to the discovery of oscillations in intracellular Ca^{2+} . Oscillations in cytosolic Ca^{2+} have since been found in a variety of cells where they can arise spontaneously, or after stimulation by hormones or neurotransmitters. Their period can

range from seconds to minutes, depending on the cell type¹⁵. These oscillations are often accompanied by propagation of intracellular or intercellular Ca^{2+} waves. The significance of Ca^{2+} oscillations and waves stems from the crucial importance of this ion in the control of many key cellular processes¹⁵.

In cells that use Ca^{2+} as second messenger, binding of an external signal to a cell-membrane receptor activates phospholipase C (PLC), which in turn synthesizes inositol 1,4,5-trisphosphate (InsP_3). This metabolite then binds to an InsP_3 receptor located on the membrane of internal Ca^{2+} stores (endoplasmic or sarcoplasmic reticulum) and thereby triggers the release of Ca^{2+} into the cytoplasm of the cell¹⁵. A conspicuous feature of Ca^{2+} release is that it is self-amplified: cytoplasmic Ca^{2+} triggers the release of Ca^{2+} from intracellular stores, a process known as Ca^{2+} -induced Ca^{2+} release (CICR).

A first model for cytosolic Ca^{2+} oscillations was based¹⁶ on the activation of PLC by Ca^{2+} . Although this positive feedback regulation has been observed in some cell types, it seems that a more general feedback process underlying oscillations is CICR itself. The effect of CICR positive feedback is antagonized by several regulatory processes (see below). A simple two-variable model for signal-induced Ca^{2+} oscillations based on CICR accounts¹⁷ for sustained oscillations of cytosolic Ca^{2+} . These oscillations occur between two critical values of the stimulus intensity, for example, two critical levels of the hormonal signal (see figure in Box 1). Below the lower critical value, a low steady-state level of cytosolic Ca^{2+} is established; above the larger critical value, the system evolves towards a higher, stable steady-state level of cytosolic Ca^{2+} . The model predicts that the frequency of Ca^{2+} oscillations rises with the degree of stimulation, as observed experimentally. In this minimal model the level of intracellular InsP_3 is treated as a control parameter reflecting the degree of external stimulation. More complex models for Ca^{2+} oscillations are based on more detailed descriptions of InsP_3 -receptor kinetics¹⁸, but still attribute to CICR a primary role in the origin of repetitive Ca^{2+} spiking.

Mathematical models for Ca^{2+} signalling have developed in two additional directions. First, waves of intra- or intercellular Ca^{2+} have been modelled by incorporating the diffusion of cytosolic Ca^{2+} or the passage of Ca^{2+} or InsP_3 from cell to cell through gap junctions^{19–22}. Although most models for Ca^{2+} waves are deterministic, stochastic simulations were used to clarify the nature of local increases of cytosolic Ca^{2+} known as blips or puffs, which are thought to trigger the onset of waves^{15,23}. Second, computational biology enables one to probe mechanisms for encoding Ca^{2+} spikes in terms of their frequency. A variety of physiological responses are controlled by the frequency and waveform of Ca^{2+} oscillations, such as gene expression during development²⁴. Among the processes that could underlie such frequency encoding are protein (de)phosphorylation by a Ca^{2+} -dependent kinase (phosphatase)¹⁷, or the Ca^{2+} -dependence of calmodulin-kinase II²⁵. A recent study combining experimental and modelling approaches showed the possibility of frequency encoding of Ca^{2+} spikes by interplay with cAMP signalling²⁶.

Pulsatile signalling in intercellular communication

Although intracellular information can be encoded in the frequency of signal-induced Ca^{2+} spikes, some extracellular signals can themselves be produced in a pulsatile manner. Examples of pulsatile intercellular communication include episodic hormone secretion and pulsatile signals of cAMP in the slime mould *Dictyostelium discoideum*. After starvation, these amoebae undergo a transition from a unicellular to a multicellular phase of their life cycle. By a chemotactic response to cAMP signals, as many as 10^5 amoebae collect around cells behaving as aggregation centres. These centres release cAMP with a period of about 5 minutes; surrounding cells relay the chemotactic signal towards the periphery of the aggregation field. Relay and oscillations of cAMP result in the formation of concentric or spiral waves of aggregating cells²⁷.

Models help to clarify the mechanism of cAMP oscillations in *Dictyostelium*^{28,29}, which involves both positive and negative feedback.

Binding of extracellular cAMP to a cell-surface receptor leads to the activation of adenylate cyclase, which catalyses the synthesis of intracellular cAMP. Transport of cAMP into the extracellular medium creates a positive feedback loop that drives a rapid rise in cAMP synthesis (phase 1 in Fig. 1a). For sustained oscillations to occur, this rise in cAMP must be self-limiting, so that cAMP first levels off before decreasing to its minimum level (phase 2). Models confirm²⁸ that negative feedback attributable to cAMP-induced receptor desensitization through reversible phosphorylation can have such a role in limiting self-amplification. Once the levels of intra- and extracellular cAMP are sufficiently low, dephosphorylation can resensitize the receptor. The ensuing build-up of extracellular cAMP (phase 3) progressively brings it to the threshold above which self-amplification triggers a new pulse.

Numerical simulations indicate that relay of cAMP pulses represents a different mode of dynamic behaviour, closely related to oscillations. Just before autonomous oscillations break out, cells in a stable steady state can amplify suprathreshold variations in extracellular cAMP in a pulsatory manner^{28,29}. Thus, relay and oscillations of cAMP are produced by a unique mechanism in adjacent domains in parameter space. The two types of dynamic behaviour are analogous to the excitable or pacemaker behaviour of nerve cells.

Theoretical models shed light on additional aspects of pulsatile cAMP signalling in *Dictyostelium*. First, like Ca^{2+} spikes, cAMP pulses are frequency encoded. Only pulses delivered at 5-min intervals are capable of accelerating slime-mould development after starvation. Simulations indicate that frequency encoding is based on reversible receptor desensitization²⁸. The kinetics of receptor resensitization dictates the interval between successive pulses required for a maximum relay response.

Second, cAMP oscillations in *Dictyostelium* provide a prototype for the ontogenesis of biological rhythms. The amoebae become capable of relaying extracellular cAMP pulses only a few hours after the beginning of starvation, before acquiring the property of autonomous oscillations. Models show that these developmental transitions can be brought about by the continuous increase in certain biochemical parameters such as the activities of adenylate cyclase or phosphodiesterase, the enzyme that degrades cAMP. In parameter space these biochemical changes define a developmental path that successively crosses domains corresponding to different types of dynamic behaviour, from no relay to relay, and finally to oscillations³.

Third, models are being used to probe the mechanisms underlying the formation of concentric or spiral waves of cAMP responsible for the spatiotemporal patterns observed during aggregation. Among the factors shown to be important in the transition between the two types of waves are extracellular phosphodiesterase activity³⁰ and desynchronization of cells that follow the developmental path after starvation³¹. Models based on the same feedback mechanism also account for the propagation of planar and scroll waves within the multicellular slug formed by the amoebae after aggregation³².

Pulsatile cAMP signalling in *Dictyostelium* is closely related with pulsatile hormone secretion in higher organisms. It is now clear that most hormones are secreted in a pulsatile rather than continuous manner³³ and that the temporal pattern of a hormone is often as important as its concentration in the blood³⁴. The best examples of pulsatile hormone signals are those of gonadotropin-releasing hormone (GnRH) secreted by the hypothalamus with a periodicity of 1 h in humans and rhesus monkey³⁵, growth hormone (GH) secreted with a period of 3 to 5 h³⁶, and insulin secreted by pancreatic β -cells with a period close to 13 min in humans³⁷. In the cases of GnRH and GH — the effect is less clear-cut for insulin — the frequency of the pulses governs the physiological efficacy of hormone stimulation^{35,36}.

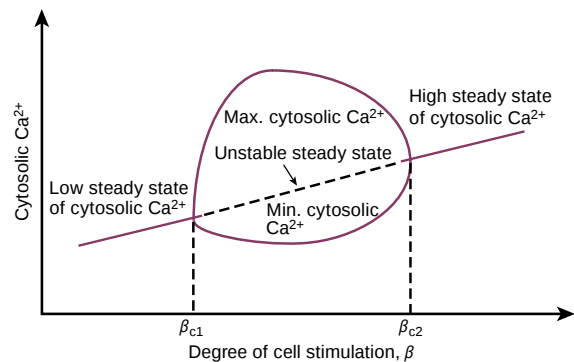
A general model for a two-state receptor subjected to periodic ligand variations shows that frequency encoding of hormone pulses may rely on reversible desensitization in target cells, as it does for cAMP pulses in *Dictyostelium*^{38,39}. The mechanism of GnRH pulsatility is still unknown and provides a challenge for both experiments

Box 1

Bifurcation diagrams

Plotting the maximum and minimum of the oscillations of a given variable as a function of a control parameter allows the construction of a 'bifurcation diagram' showing where the system changes its dynamical properties. A common type of bifurcation diagram is depicted in the figure opposite for the case of intracellular Ca^{2+} oscillations. Below a critical value, the system reaches a stable steady state, while at the critical value the steady state becomes unstable and a stable limit cycle begins to grow, surrounding the unstable steady state (the critical value at which the limit cycle appears corresponds to a Hopf bifurcation). The amplitude of the limit cycle increases and passes through a maximum as the value of the control parameter increases. Finally, above a second, higher critical value, sustained oscillations disappear and the system again evolves towards a new stable steady state. This bifurcation diagram illustrates an important property of sustained oscillations, namely that they occur within a certain parameter range often bounded by two critical values.

The scheme in figure opposite represents the simplest type of bifurcation diagram for the onset of sustained oscillations. More complex bifurcation diagrams are obtained when multiple attractors (that is, stable steady states or stable oscillations) coexist in a certain range of parameter values. Thus a stable steady state and a stable limit cycle, corresponding to sustained oscillations, may coexist, separated by an unstable limit cycle. Such a situation is referred to as hard excitation. Other modes of attractor multiplicity include the coexistence between two stable steady states (bistability) or between two stable limit cycles (birhythmicity). Each of the stable attractors possesses its basin of attraction, which includes all initial conditions, in phase space, from which the system evolves towards this particular attractor.



Box 1 Figure Schematic bifurcation diagram showing the domain and amplitude of intracellular Ca^{2+} oscillations as a function of the degree of external stimulation β , which is used as control parameter. Sustained Ca^{2+} oscillations occur in a range of stimulation between the two critical β -values denoted β_{c1} and β_{c2} . The maximum and minimum of cytosolic Ca^{2+} oscillations are plotted as a function of β in this range, in which the dashed line refers to the unstable steady state. On the left and right sides of the oscillatory domain, the system evolves to a stable steady state (solid line) corresponding to a low and high level of cytosolic Ca^{2+} , respectively. In the situation described, a unique steady state corresponds to a given value of β ; the precise value of the steady state depends on β . The bifurcation diagram is obtained in a two-variable model¹⁷ for Ca^{2+} oscillations based on Ca^{2+} -induced Ca^{2+} release (see ref. 17 for a nonschematic version of the diagram).

and theory. The basis of pulsatile GH secretion has been studied by a modelling approach⁴⁰. In β -cells, pulsatile insulin release could originate from insulin feedback on glucose transport into the cells⁴¹ or from oscillatory membrane activity driven by glycolytic oscillations³⁷. Together with these metabolic oscillations, membrane-potential bursting and Ca^{2+} oscillations in β -cells illustrate the multiplicity of rhythms that can be encountered in a given cell type.

Circadian rhythms

The most ubiquitous biological rhythms are those that occur with a period close to 24 h and that allow organisms to adapt to periodic variations in the terrestrial environment. Experimental advances during the past decade have clarified the molecular bases of these circadian rhythms, first in *Drosophila* and *Neurospora*, and more recently in cyanobacteria, plants and mammals^{42–44}. In all cases investigated so far, it appears that circadian rhythms originate from the negative feedback exerted by a protein on the expression of its gene⁴⁵.

Before details on the molecular mechanism of circadian rhythms began to be uncovered, theoretical models borrowed from physics were used to investigate their dynamic properties. The relative simplicity of these models explains why their use continues to this day. Thus the Van der Pol equations, derived for an electrical oscillator, served for modelling the response of human circadian oscillations to light⁴⁶ and to account for experimental observations on increased fitness due to resonance of the circadian clock with the external light–dark cycle in cyanobacteria⁴⁷. The earliest model predicting oscillations due to negative feedback on gene expression was proposed by Goodwin⁴⁸, at a time when the part played by such a regulatory mechanism in the origin of circadian rhythms was not yet known. Models of this type are still being used in studies of circadian oscillations, for example in *Neurospora*⁴⁹.

Molecular models for circadian rhythms were proposed⁵⁰ initially for circadian oscillations of the period (PER) protein and its mRNA in

Drosophila, the first organism for which detailed information on the oscillatory mechanism became available⁴⁵ (the PER protein behaves as a transcriptional regulator capable of influencing the expression of a variety of genes besides its own gene, *per*). The case of circadian rhythms in *Drosophila* illustrates how the need to incorporate experimental advances leads to a progressive increase in the complexity of theoretical models. A first model⁵⁰ governed by a set of five kinetic equations is shown in Fig. 2a; it is based on the negative control exerted by the PER protein on the expression of *per*. Numerical simulations show that for appropriate parameter values, the steady state becomes unstable and limit-cycle oscillations appear (Fig. 1b, d).

This early model did not account for the effect of light on the circadian system. Experiments subsequently showed that a second protein, timeless or TIM, forms a complex with PER, and that light acts by inducing TIM degradation⁴³. An extended, ten-variable model was then proposed⁵¹, in which the negative regulation is exerted by the PER–TIM complex (Fig. 2b). This model produces essentially the same result, sustained oscillations in continuous darkness. In addition, it accounts for the behaviour of mutants and explicitly incorporates the effect of light on the TIM degradation rate. Thereby the model can account for the entrainment of the oscillations by light–dark cycles and for the phase shifts induced by light pulses⁵¹. A closely related model incorporating the formation of a PER–TIM complex has been proposed for *Drosophila* circadian rhythms⁵².

Theoretical models for circadian rhythms in *Drosophila* bear on the mechanism of circadian oscillations in mammals, where homologues of the *per* gene exist and negative autoregulation of gene expression is also found⁴⁴. However, in mammals, the role of TIM as a partner for PER is played by the cryptochrome (CRY) protein, and light acts by inducing gene expression rather than protein degradation as in *Drosophila*⁴⁴. A further analogy between *Drosophila* and mammals is that the negative feedback on gene expression is indirect: the PER–TIM or PER–CRY complexes exert their repressive effect by binding to a complex of two

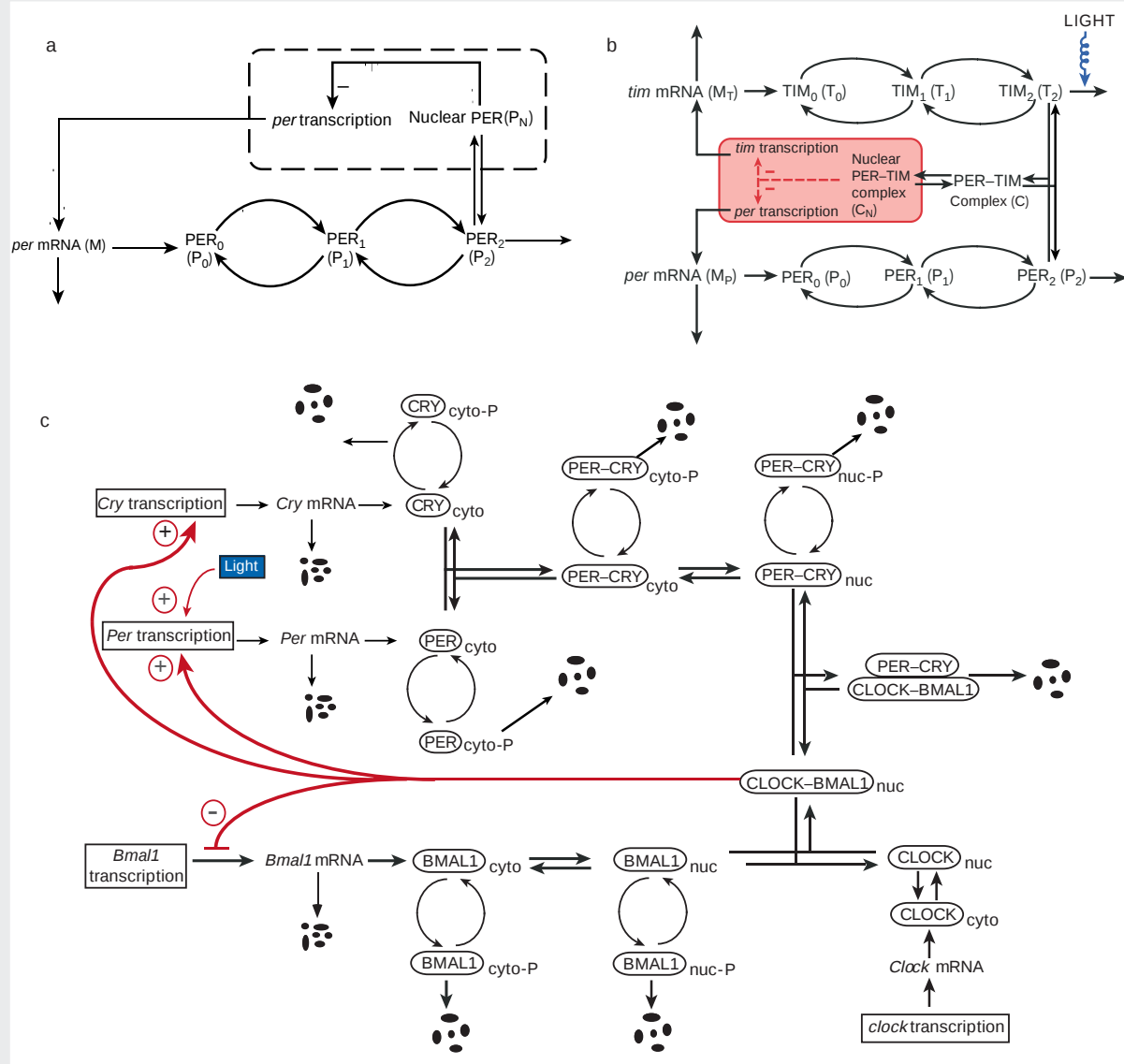


Figure 2 Molecular models of increasing complexity considered for circadian oscillations. **a**, Model for circadian oscillations in *Drosophila* based on negative autoregulation of the *per* gene by its protein product PER^{3,50}. The model incorporates gene transcription into *per* mRNA, transport of *per* mRNA into the cytosol as well as mRNA degradation, synthesis of the PER protein at a rate proportional to the *per* mRNA level, reversible phosphorylation and degradation of PER, as well as transport of PER into the nucleus where it represses the transcription of the *per* gene. The model is described by a set of five kinetic equations^{3,50}. **b**, Model for circadian oscillations in *Drosophila* incorporating the formation of a complex between the PER and TIM proteins⁵¹. The model is described by a set of ten kinetic equations⁵¹. **c**, Model for circadian oscillations in mammals incorporating indirect, negative autoregulation

of the *Per* and *Cry* genes through binding of the PER-CRY dimer to the complex formed between the two activating proteins CLOCK and BMAL1. Also considered is the negative feedback exerted by the latter proteins on the expression of their genes. Synthesis, reversible phosphorylation, and degradation of the various proteins are taken into account. The model is described by a set of 16 kinetic equations⁵⁸. For appropriate parameter values, all three models admit sustained circadian oscillations in conditions corresponding to continuous darkness. The effect of light is taken into account in the models in **b** and **c** by incorporating light-induced TIM degradation or light-induced *Per* expression, respectively. Further extensions of the model shown in **c** for the mammalian clock are needed to incorporate the recently discovered role of the *Rev-erba* and *Dec* genes^{55,89}.

proteins, CLOCK-CYC or CLOCK-BMAL1 in the fly⁵³ and in mammals⁵⁴, respectively. These proteins activate *per* and *tim* (or *cry*) gene expression. Thus negative feedback occurs by counteracting the effect of gene activators. Additional feedback loops are present, such as the negative feedback exerted by CLOCK or BMAL1 on the expression of their genes. These controls, which are mediated by other gene products^{44,55}, are also removed upon formation of the complex with the PER-TIM or PER-CRY dimers^{53,54}.

What are the dynamical consequences of these additional regulatory loops and of the indirect path of the negative feedback on gene

expression? Addressing these issues requires further extension of the model. Such an extended model has been proposed for *Drosophila*^{56,57} and is currently being studied for mammals⁵⁸. The model for the circadian clock mechanism in mammals is schematized in Fig. 2c. The presence of additional mRNA and protein species, as well as of multiple complexes formed between the various clock proteins, complicates the model. The time evolution of this extended model is governed by a system of 16 kinetic equations. Sustained or damped oscillations can occur in this model for parameter values corresponding to continuous darkness. As observed in the experiments on the

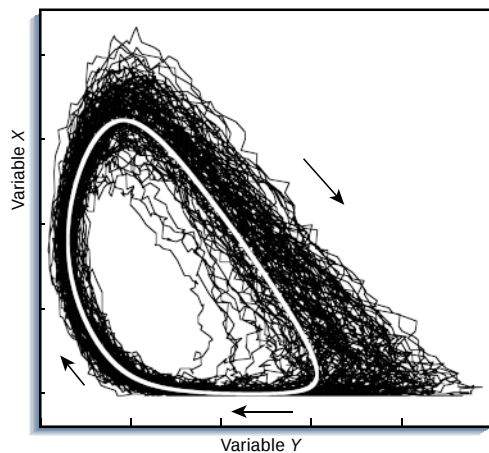


Figure 3 Effect of molecular noise on circadian oscillations. Stochastic simulations of the model of Fig. 2a yield oscillations that correspond, in the phase plane (X , Y), to the evolution to a noisy limit cycle. The latter takes the form of a cloud of points surrounding the deterministic limit cycle (white solid curve) obtained in the absence of molecular noise (data redrawn from Fig. 3a in ref. 64). Variables X (*per*mRNA) and Y (nuclear form of the PER protein) are expressed as numbers of molecules or concentrations in stochastic and deterministic simulations, respectively. In the stochastic simulations illustrated, X and Y vary in the range between 0 and 200 molecules and 20 and 800 molecules, respectively⁶⁴. Arrows indicate the direction of movement along the deterministic or stochastic limit cycle.

mammalian clock, *Bmal1* mRNA oscillates in opposite phase with respect to *Per* and *Cry* mRNAs⁴⁴. Entrainment by the external light–dark cycle can be captured by the model if it incorporates the light-induced increase in the rate of *Per* expression. Numerical simulations show that, upon entrainment, a slight change in parameters, such as the maximum rate of PER phosphorylation, suffices to shift the peak in *Per* mRNA with respect to the onset of the light phase by several hours. This lability could explain why the phase of circadian oscillations in mammals varies in peripheral tissues with respect to the phase of the central pacemaker located in the suprachiasmatic nuclei within the hypothalamus⁴⁴.

The results obtained with the model for the mammalian circadian clock provide cues for circadian rhythm sleep disorders in humans⁵⁹. Thus permanent phase shifts in light–dark conditions could account for the familial advanced sleep phase syndrome that has been attributed to PER hypophosphorylation⁶⁰, and for the delayed sleep phase syndrome, which is also related to PER⁶¹. For some parameter values the model fails to allow entrainment by 24-h light–dark cycles. This result could account for the non-24-h sleep–wake syndrome in which the phase of the sleep–wake pattern varies continuously with respect to the light–dark cycle, that is, the patient free-runs in light–dark conditions⁵⁹.

Computational biology can provide surprisingly counterintuitive insights. A case in point is the puzzling observation that circadian rhythms in continuous darkness can sometimes be suppressed by a single pulse of light and restored by a second such pulse. Winfree² proposed the first theoretical explanation for this long-term suppression. He hypothesized that the limit cycle in each oscillating cell surrounds an unstable steady state. The light pulse would act as a critical perturbation that would bring the clock to the singularity, that is, the steady state. Because the steady state is unstable, each cell would eventually return to the limit cycle, but the population would be spread out over the entire cycle so that the cells would be desynchronized and no global rhythm would be seen.

An alternative explanation is based on the coexistence of sustained oscillations with a stable steady state. Such coexistence has been observed⁶², albeit in a restricted domain in parameter space, in the

model for circadian rhythms in *Drosophila* based on negative autoregulation by the PER–TIM complex (Fig. 2b). In such a situation, the effect of the light pulse is to bring the clock mechanism into the basin of attraction of the stable steady state in each oscillating cell, so that the rhythm is suppressed. A second light pulse then brings the system back to the limit cycle's basin of attraction corresponding to circadian oscillations⁶². Without a computational model it is impossible to predict the coexistence between a stable steady state and a stable rhythm.

I have discussed only deterministic models for cellular rhythms so far. Do the models remain valid when the numbers of molecules involved are small, as may occur in cellular conditions? In the presence of small amounts of mRNA or protein molecules, the effect of molecular noise on circadian rhythms may become significant and may compromise the emergence of coherent periodic oscillations⁶³. The way to assess the influence of molecular noise is to resort to stochastic simulations (see review in this issue by Rao and colleagues, pages 231–237). In applying this approach to the models for circadian rhythms schematized in Fig. 2a, b, we must first break down the different reactions into elementary steps.

The temporal dynamics of the system is determined numerically by allowing the various reaction steps to occur randomly, with a frequency measured by their probability of occurrence. These stochastic simulations show that the dynamic behaviour predicted by the corresponding deterministic equations remains valid as long as the maximum numbers of mRNA and protein molecules involved in the circadian clock mechanism are of the order of a few tens and hundreds, respectively⁶⁴. The larger the numbers of molecules, the smaller the noise due to random fluctuations. In the presence of molecular noise, the trajectory in the phase space transforms into a cloud of points surrounding the deterministic limit cycle (Fig. 3). Stochastic simulations confirm the existence of bifurcation values of the control parameters bounding a domain in which sustained oscillations occur⁶⁴. Only when the maximum numbers of molecules of mRNA and protein become smaller than a few tens does noise begin to obliterate the circadian rhythm. Mechanisms enhancing resistance to noise in genetic oscillators have been investigated in a recent theoretical study⁶⁵.

From simple to complex oscillatory behaviour

Computational biology clarifies the mechanisms responsible for the transition from simple to complex oscillatory phenomena in biochemical and cellular systems^{3,66}. Bursting represents one type of complex oscillations that is particularly common in neurobiology. An active phase of spike generation is followed by a quiescent phase, after which a new active phase begins. Mathematical models throw light on the conditions that generate these complex periodic oscillations⁶⁷. Chaos is a common mode of complex oscillatory behaviour that has been studied intensively in physical, chemical and biological systems^{3,68,69}. In phase space, chaotic oscillations correspond to the evolution towards a so-called strange attractor. These irregular oscillations are characterized by their sensitivity to initial conditions, which accounts for the unpredictable nature of chaotic dynamics. Yet another type of complex oscillatory behaviour involves the coexistence of multiple attractors. When a stable steady state and a stable limit cycle coexist (as in the case of suppression of circadian rhythm discussed above), they conspire to produce what is called hard excitation. Two stable limit cycles may also coexist, separated by an unstable cycle. This phenomenon, referred to as birhythmicity^{3,66}, is the oscillatory counterpart of bistability in which two stable steady states, separated by an unstable state, coexist. Birhythmicity was predicted by numerical simulations before being observed experimentally³.

The study of models indicates the existence of two main routes to complex oscillatory phenomena. The first relies on forcing a system that displays simple periodic oscillations by a periodic input⁶⁹. In an appropriate range of input frequency and amplitude, one can often observe the transition from simple to complex oscillatory behaviour such as bursting and chaos. For other frequencies and amplitudes of

the forcing, entrainment or quasiperiodic oscillations occur. Circadian rhythms are subjected to periodic forcing naturally by light–dark cycles, and numerical simulations of a model for the circadian clock indicate that entrainment, quasiperiodic oscillations and chaos may occur, depending on the magnitude of the periodic changes induced by the light–dark cycle in the light-sensitive parameter⁶⁶. The waveform of the forcing is also important, as the domain of entrainment enlarges at the expense of chaos when the input transforms from a square wave into a sinusoidal forcing⁶⁶.

Complex oscillations can also occur in autonomous systems that operate in a constant environment. The study of models for a variety of cellular oscillations shows that complex oscillatory phenomena may arise through the interplay between several instability-generating mechanisms, each of which is capable of producing sustained oscillations^{3,66}. The case of Ca^{2+} signalling is particularly revealing, because of the multiplicity of feedback mechanisms that could potentially be involved in the onset of oscillations. Thus, among the many nonlinear processes that could take part in an instability-generating loop are: (1) Ca^{2+} -induced Ca^{2+} release; (2) desensitization of the InsP_3 receptor; (3) bell-shaped dependence of the InsP_3 receptor on Ca^{2+} , which reflects its activation and inhibition at different Ca^{2+} levels; (4) capacitative Ca^{2+} entry; (5) PLC or/and InsP_3 3-kinase activation by Ca^{2+} ; (6) control of Ca^{2+} by mitochondria; (7) G-protein regulation by Ca^{2+} ; and (8) coupling of the membrane potential to cytosolic Ca^{2+} . Several models in which at least two of these regulatory processes are coupled were shown to admit birhythmicity, bursting or chaotic oscillations^{22,66,70,71}.

Concluding remarks

Given the rapid accumulation of new data on gene, protein and cellular networks, it is increasingly clear that computational biology will be crucial in making sense of the puzzle of cellular regulatory interactions. Models and simulations are particularly valuable for exploring the dynamic phenomena associated with these regulations. Such an approach has long been applied to the study of biological rhythms, from the periodic activity of nerve and cardiac cells to population oscillations in ecology. I have focused here on the computational biology of some of the main oscillatory phenomena that arise at the cellular level. Additional examples of cellular oscillatory processes that have been studied by means of theoretical models abound. A most important one is the eukaryotic cell cycle. Models indicate that mitosis in amphibian embryonic cells is driven by a limit-cycle oscillator that produces the repetitive activation of the cyclin-dependent kinase cdk1 (refs 72, 73). The interplay between oscillations and bistability has been addressed in detailed molecular models for the cell cycles of yeast and somatic cells, which are more complex owing to the existence of checkpoints^{74,75}.

At the genetic level, models show that regulatory interactions between genes can result in multiple steady states or oscillations. The two types of phenomena have recently been demonstrated in synthetic genetic networks reconstituted in bacteria (refs 76, 77; and see review in this issue by Hasty and colleagues, pages 224–230). Circadian rhythms are a fertile field for applying computational biology to the study of oscillations associated with genetic regulation. Also related to the regulation of gene expression are the oscillations in the activity of the tumour suppressor p53, which have been studied both experimentally and by means of a model⁷⁸. A segmentation clock involving Notch signalling is responsible for periodic somite formation⁷⁹. Oscillatory nucleocytoplasmic shuttling of the Msn2 transcription factor, with a period of several minutes, has recently been observed in yeast and studied theoretically⁸⁰.

Glycolytic oscillations represent the prototype of periodic behaviour associated with enzyme regulation^{3,13,14}. Early models for glycolytic oscillations were centred around the reaction catalysed by phosphofructokinase, and took into account the allosteric and regulatory properties of this product-activated enzyme^{3,14,81}. More detailed models^{82–84} take into account the full set of glycolytic enzyme

reactions. In these models, the primary role played by phosphofructokinase in the instability-generating mechanism is somewhat blurred. This example illustrates the two main approaches followed in computational biology. The first is based on minimal models—a complex system is decomposed into simpler modules⁸⁵, each of which can be modelled by simple equations. Once these are understood, they are assembled into increasingly complex networks that can exhibit collective properties not apparent in the modules' behaviour. The second relies on large-scale models that aim at incorporating from the outset all known details about the variables and processes of interest. This approach may someday lead to the construction of an electronic cell *in silico*, although that day remains far off. With models as with maps, I believe that an intermediate scale will often prove most fruitful.

The comparison of theoretical predictions with experimental results calls for more quantitative data in molecular and cell biology⁸⁶. The advent of new tools should facilitate the collection of more quantitative data on the dynamics of cellular processes. Such data will complement qualitative studies on the nature of interactions in cellular regulatory networks.

Clarification of the molecular mechanisms underlying oscillations is but one application of computational biology to the study of cellular rhythms. As discussed in this review, models are also used to address the function of these rhythms, which is often related to their frequency encoding, and a variety of related phenomena such as propagating waves and complex oscillations. The link between intracellular oscillations and the propagation of intra- or intercellular waves is well illustrated by Ca^{2+} signalling in many cell types and by cAMP signalling in *Dictyostelium* amoebae. Recent observations on the occurrence of intracellular waves in activated leukocytes⁸⁷ provide a challenge for modelling studies. This spatiotemporal phenomenon seems to be linked with the occurrence of metabolic oscillations, the nature of which are unclear. The modelling approach has been applied to account for the transition from simple to complex oscillatory behaviour in the peroxidase reaction⁶⁸ and in the Ca^{2+} signalling system^{22,66,70,71}. The observation of a transition to chaos in the glycolytic cycle in yeast cell cultures⁸⁸ remains to be studied in a similar manner.

Models for cellular rhythms illustrate the roles and advantages of computational biology. First and foremost, modelling takes over when pure intuition reaches its limits. This situation commonly arises when studying cellular processes that involve a large number of variables coupled through multiple regulatory interactions. Here one cannot make reliable predictions on the basis of verbal reasoning. But mathematical models can show the precise parameter ranges that give rise to sustained oscillations. Models also help clarify the molecular mechanisms of these oscillations. Indeed, simulations allow rapid determination of the qualitative and quantitative effects of each parameter, and thereby can help to identify key parameters that have the most profound effect on the system's dynamics. Testing various models permits swift exploration of different mechanisms over a large range of conditions. One of the main roles of models will be to provide a unified conceptual framework to account for experimental observations and to generate testable predictions.

From a more global perspective, which represents one of the strengths of the theoretical approach, the common mathematical structure of models underlines the links between similar dynamic phenomena occurring in widely different biological settings, from genetic to metabolic and neural networks, and from cell to animal populations. □

doi:10.1038/nature01259

1. Fessard, A. *Propriétés Rythmiques de la Matière Vivante* (Hermann, Paris, 1936).
2. Winfree, A. T. *The Geometry of Biological Time* 2nd edn (Springer, New York, 2001).
3. Goldbeter, A. *Biochemical Oscillations and Cellular Rhythms: The Molecular Bases of Periodic and Chaotic Behaviour* (Cambridge Univ. Press, Cambridge, 1996).
4. Volterra, V. Fluctuations in the abundance of a species considered mathematically. *Nature* **118**, 558–560 (1926).

5. Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane currents and its application to conduction and excitation in nerve. *J. Physiol. (Lond.)* **117**, 500–544 (1952).
6. Koch, C. & Segev, I. (eds) *Methods in Neuronal Modeling. From Synapses to Networks* 2nd edn (MIT Press, Cambridge, MA, 1998).
7. Keener, J. P. & Sneyd, J. *Mathematical Physiology* (Springer, New York, 1998).
8. Noble, D. Modeling the heart—from genes to cells to the whole organ. *Science* **295**, 1678–1682 (2002).
9. Nicolis, G. & Prigogine, I. *Self-Organization in Nonequilibrium Systems. From Dissipative Structures to Order through Fluctuations* (Wiley, New York, 1977).
10. Thomas, R. & d'Ari, R. *Biological Feedback* (CRC Press, Boca Raton, FL, 1990).
11. Ferrell, J. E. Jr Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. *Curr. Opin. Cell Biol.* **14**, 140–148 (2002).
12. Doedel, E. J. AUTO: A program for the automatic bifurcation analysis of autonomous systems. *Cong. Numer.* **30**, 265–284 (1981). (Available at <http://ftp.cs.concordia.ca/pub/doedel/auto/>.)
13. Hess, B. & Boiteux, A. Oscillatory phenomena in biochemistry. *Annu. Rev. Biochem.* **40**, 237–258 (1971).
14. Goldbeter, A. & Caplan, S. R. Oscillatory enzymes. *Annu. Rev. Biophys. Bioeng.* **5**, 449–476 (1976).
15. Berridge, M. J. Elementary and global aspects of calcium signalling. *J. Physiol. (Lond.)* **499**, 291–306 (1997).
16. Meyer, T. & Stryer, L. Molecular model for receptor-stimulated calcium spiking. *Proc. Natl Acad. Sci. USA* **85**, 5051–5055 (1988).
17. Goldbeter, A., Dupont, G. & Berridge, M. J. Minimal model for signal-induced Ca^{2+} oscillations and for their frequency encoding through protein phosphorylation. *Proc. Natl Acad. Sci. USA* **87**, 1461–1465 (1990).
18. De Young, G. W. & Keizer, J. A single-pool inositol 1,4,5-trisphosphate-receptor-based model for agonist-stimulated oscillations in Ca^{2+} concentration. *Proc. Natl Acad. Sci. USA* **89**, 9895–9899 (1992).
19. Dupont, G. & Goldbeter, A. Properties of intracellular Ca^{2+} waves generated by a model based on Ca^{2+} -induced Ca^{2+} release. *Biophys. J.* **67**, 2191–2204 (1994).
20. Sneyd, J., Charles, A. C. & Sanderson, M. J. A model for the propagation of intercellular calcium waves. *Am. J. Physiol.* **266**, C293–C302 (1994).
21. Dupont, G. *et al.* Mechanism of receptor-oriented intercellular calcium wave propagation in hepatocytes. *FASEB J.* **14**, 279–289 (2000).
22. Schuster, S., Marhl, M. & Höfer, T. Modelling of simple and complex calcium oscillations. From single-cell responses to intercellular signalling. *Eur. J. Biochem.* **269**, 1333–1355 (2002).
23. Swillens, S., Dupont, G., Combettes, L. & Champeil, P. From calcium blips to calcium puffs: theoretical analysis of the requirements for interchannel communication. *Proc. Natl Acad. Sci. USA* **96**, 13750–13755 (1999).
24. Spitzer, N. C., Lautermilch, N. J., Smith, R. D. & Gomez, T. M. Coding of neuronal differentiation by calcium transients. *BioEssays* **22**, 811–817 (2000).
25. De Koninck, P. & Schulman, H. Sensitivity of CaM kinase II to the frequency of Ca^{2+} oscillations. *Science* **279**, 227–230 (1998).
26. Gorbunova, Y. V. & Spitzer, N. C. Dynamic interactions of cyclic AMP transients and spontaneous Ca^{2+} spikes. *Nature* **418**, 93–96 (2002).
27. Dormann, D., Kim, J. Y., Devreotes, P. N. & Weijer, C. J. cAMP receptor affinity controls wave dynamics, geometry and morphogenesis in *Dictyostelium*. *J. Cell Sci.* **114**, 2513–2523 (2001).
28. Martiel, J. L. & Goldbeter, A. A model based on receptor desensitization for cyclic AMP signaling in *Dictyostelium* cells. *Biophys. J.* **52**, 807–828 (1987).
29. Tang, Y. & Othmer, H. G. Excitation, oscillations and wave propagation in a G-protein-based model of signal transduction in *Dictyostelium discoideum*. *Phil. Trans. R. Soc. Lond. B* **349**, 179–195 (1995).
30. Palsson, E. & Cox, E. C. Origin and evolution of circular waves and spirals in *Dictyostelium discoideum* territories. *Proc. Natl Acad. Sci. USA* **93**, 1151–1155 (1996).
31. Lauzeral, J., Halloy, J. & Goldbeter, A. Desynchronization of cells on the developmental path triggers the formation of spiral waves of cAMP during *Dictyostelium* aggregation. *Proc. Natl Acad. Sci. USA* **94**, 9153–9158 (1997).
32. Bretschneider, T., Siegert, F. & Weijer, C. J. Three-dimensional scroll waves of cAMP could direct cell movement and gene expression in *Dictyostelium* slugs. *Proc. Natl Acad. Sci. USA* **92**, 4387–4391 (1995).
33. Chadwick, D. J. & Goode, J. A. (eds) *Mechanisms and Biological Significance of Pulsatile Hormone Secretion* (Novartis Found. Symp. **227**) (Wiley, Chichester, 2000).
34. Knobil, E. Patterns of hormone signals and hormone action. *New Engl. J. Med.* **305**, 1582–1583 (1981).
35. Belchetz, P. E., Plant, T. M., Nakai, Y., Keogh, E. J. & Knobil, E. Hypophysal responses to continuous and intermittent delivery of hypothalamic gonadotropin-releasing hormone. *Science* **202**, 631–633 (1978).
36. Hindmarsh, P. C., Stanhope, R., Preece, M. A. & Brook, C. G. D. Frequency of administration of growth hormone—an important factor in determining growth response to exogenous growth hormone. *Horm. Res.* **33**(Suppl. 4), 83–89 (1990).
37. Tornheim, K. Are metabolic oscillations responsible for normal oscillatory insulin secretion? *Diabetes* **46**, 1375–1380 (1997).
38. Li, Y. X. & Goldbeter, A. Frequency specificity in intercellular communication: the influence of patterns of periodic signalling on target cell responsiveness. *Biophys. J.* **55**, 125–145 (1989).
39. Goldbeter, A., Dupont, G. & Halloy, J. The frequency encoding of pulsatility. *Novartis Found. Symp.* **227**, 19–36 (2000).
40. Wagner, C., Caplan, S. R. & Tannenbaum, G. S. Genesis of the ultradian rhythm of GH secretion: a new model unifying experimental observations in rats. *Am. J. Physiol.* **275**, E1046–E1054 (1998).
41. Maki, L. W. & Keizer, J. Mathematical analysis of a proposed mechanism for oscillatory insulin secretion in perfused HIT-15 cells. *Bull. Math. Biol.* **57**, 569–591 (1995).
42. Dunlap, J. C. Molecular bases for circadian clocks. *Cell* **96**, 271–290 (1999).
43. Young, M. W. & Kay, S. A. Time zones: a comparative genetics of circadian clocks. *Nature Rev. Genet.* **2**, 702–715 (2001).
44. Reppert, S. M. & Weaver, D. R. Coordination of circadian timing in mammals. *Nature* **418**, 935–941 (2002).
45. Hardin, P. E., Hall, J. C. & Rosbash, M. Feedback of the *Drosophila* period gene product on circadian cycling of its messenger RNA levels. *Nature* **343**, 536–540 (1990).
46. Kronauer, R. E., Forger, D. B. & Jewett, M. E. Quantifying human circadian pacemaker response to brief, extended, and repeated light stimuli over the phototopic range. *J. Biol. Rhythms* **14**, 500–515 (1999).
47. Gonze, D., Roussel, M. & Goldbeter, A. A model for the enhancement of fitness in cyanobacteria based on resonance of a circadian oscillator with the external light-dark cycle. *J. Theor. Biol.* **214**, 577–597 (2002).
48. Goodwin, B. C. Oscillatory behavior in enzymatic control processes. *Adv. Enzyme Regul.* **3**, 425–438 (1965).
49. Ruoff, P., Vinsjevsk, M., Monnerjahn, C. & Rensing, L. The Goodwin model: simulating the effect of light pulses on the circadian sporulation rhythm of *Neurospora crassa*. *J. Theor. Biol.* **209**, 29–42 (2001).
50. Goldbeter, A. A model for circadian oscillations in the *Drosophila* period protein (PER). *Proc. R. Soc. Lond. B* **261**, 319–324 (1995).
51. Leloup, J. C. & Goldbeter, A. A model for circadian rhythms in *Drosophila* incorporating the formation of a complex between the PER and TIM proteins. *J. Biol. Rhythms* **13**, 70–87 (1998).
52. Tyson, J. J., Hong, C. I., Thron, C. D. & Novak, B. A simple model of circadian rhythms based on dimerization and proteolysis of PER and TIM. *Biophys. J.* **77**, 2411–2417 (1999).
53. Glossop, N. R., Lyons, L. C. & Hardin, P. E. Interlocked feedback loops within the *Drosophila* circadian oscillator. *Science* **286**, 766–768 (1999).
54. Shearman, L. P. *et al.* Interacting molecular loops in the mammalian circadian clock. *Science* **288**, 1013–1019 (2000).
55. Preitner, N. *et al.* The orphan nuclear receptor REV-ERB α controls circadian transcription within the positive limb of the mammalian circadian oscillator. *Cell* **110**, 251–260 (2002).
56. Ueda, H. R., Hagiwara, M. & Kitano, H. Robust oscillations within the interlocked feedback model of *Drosophila* circadian rhythm. *J. Theor. Biol.* **210**, 401–406 (2001).
57. Smolen, P., Baxter, D. A. & Byrne, J. H. Modeling circadian oscillations with interlocking positive and negative feedback loops. *J. Neurosci.* **21**, 6644–6656 (2001).
58. Leloup, J. C. & Goldbeter, A. Towards a detailed computational model for the mammalian circadian clock. *Proc. Natl Acad. Sci. USA* (submitted).
59. Richardson, G. S. & Malin, H. V. Circadian rhythm sleep disorders: pathophysiology and treatment. *J. Clin. Neurophysiol.* **13**, 17–31 (1996).
60. Toh, K. L. *et al.* An hPer2 phosphorylation site mutation in familial advanced sleep phase syndrome. *Science* **291**, 1040–1043 (2001).
61. Ebisawa, T. *et al.* Association of structural polymorphisms in the human *period3* gene with delayed sleep phase syndrome. *EMBO Rep.* **2**, 342–346 (2001).
62. Leloup, J. C. & Goldbeter, A. A molecular explanation for the long-term suppression of circadian rhythms by a single light pulse. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **280**, R1206–R1212 (2001).
63. Barkai, N. & Leibler, S. Circadian clocks limited by noise. *Nature* **403**, 267–268 (2000).
64. Gonze, D., Halloy, J. & Goldbeter, A. Robustness of circadian rhythms with respect to molecular noise. *Proc. Natl Acad. Sci. USA* **99**, 673–678 (2002).
65. Vilar, J. M., Kueh, H. Y., Barkai, N. & Leibler, S. Mechanisms of noise-resistance in genetic oscillators. *Proc. Natl Acad. Sci. USA* **99**, 5988–5992 (2002).
66. Goldbeter, A. *et al.* From simple to complex oscillatory behavior in metabolic and genetic control networks. *Chaos* **11**, 247–260 (2001).
67. Rinzel, J. A formal classification of bursting mechanisms in excitable systems. *Lect. Notes Biomath.* **71**, 267–281 (1987).
68. Olsen, L. F. & Degn, H. Chaos in biological systems. *Q. Rev. Biophys.* **18**, 165–225 (1985).
69. Glass, L. & Mackey, M. C. *From Clocks to Chaos: The Rhythms of Life* (Princeton Univ. Press, Princeton, 1988).
70. Shen, P. & Larter, R. Chaos in intracellular Ca^{2+} oscillations in a new model for non-excitable cells. *Cell Calcium* **17**, 225–232 (1995).
71. Kummer, U. *et al.* Switching from simple to complex oscillations in calcium signaling. *Biophys. J.* **79**, 1188–1195 (2000).
72. Goldbeter, A. A minimal cascade model for the mitotic oscillator involving cyclin and cdc2 kinase. *Proc. Natl Acad. Sci. USA* **88**, 9107–9111 (1991).
73. Novak, B. & Tyson, J. J. Numerical analysis of a comprehensive model of M-phase control in *Xenopus* oocyte extracts and intact embryos. *J. Cell Sci.* **106**, 1153–1168 (1993).
74. Tyson, J. J. & Novak, B. Regulation of the eukaryotic cell cycle: molecular antagonism, hysteresis, and irreversible transitions. *J. Theor. Biol.* **210**, 249–263.
75. Tyson, J. J., Chen, K. & Novak, B. Network dynamics and cell physiology. *Nature Rev. Mol. Cell Biol.* **2**, 908–916 (2001).
76. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–338 (2000).
77. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403**, 339–342 (2000).
78. Lev Bar-Or, R. *et al.* Generation of oscillations by the p53-Mdm2 feedback loop: a theoretical and experimental study. *Proc. Natl Acad. Sci. USA* **97**, 11250–11255 (2000).
79. Maroto, M. & Pourquie, O. A molecular clock involved in somite segmentation. *Curr. Top. Dev. Biol.* **51**, 221–248 (2001).
80. Jacquet, M., Renault, G., Lallet, S., de Mey, J. & Goldbeter, A. Oscillatory nucleocytoplasmic shuttling of the general stress response transcriptional activator Msn2 in *Saccharomyces cerevisiae*. *Nature* (submitted).
81. Boiteux, A., Goldbeter, A. & Hess, B. Control of oscillating glycolysis of yeast by stochastic, periodic, and steady source of substrate: a model and experimental study. *Proc. Natl Acad. Sci. USA* **72**, 3829–3833 (1975).
82. Termonia, Y. & Ross, J. Oscillations and control features in glycolysis: numerical analysis of a comprehensive model. *Proc. Natl Acad. Sci. USA* **78**, 2952–2956 (1981).
83. Hynne, F., Dano, S. & Sorensen, P. G. Full-scale model of glycolysis in *Saccharomyces cerevisiae*. *Biophys. Chem.* **94**, 121–163 (2001).
84. Reijenga, K. A., Westerhoff, H. V., Kholodenko, B. N. & Snoep, J. L. Control analysis for autonomously oscillating biochemical networks. *Biophys. J.* **82**, 99–108 (2002).
85. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**(Suppl.), C47–C52 (1999).
86. Koshland, D. E. Jr The era of pathway quantification. *Science* **280**, 852–853 (1998).
87. Petty, H. R. Neutrophil oscillations: temporal and spatiotemporal aspects of cell behavior. *Immunol. Rev.* **23**, 85–94 (2001).
88. Nielsen, K., Sorensen, P. G. & Hynne, F. Chaos in glycolysis. *J. Theor. Biol.* **186**, 303–306 (1997).
89. Honma, S. *et al.* *Dec1* and *Dec2* are regulators of the mammalian molecular clock. *Nature* **419**, 841–844 (2002).

Acknowledgements

I thank G. Dupont, D. Gonze, B. Jacrot, J. C. Leloup and G. Oster for discussions and helpful comments on the manuscript. This work was supported by a grant from the Fonds de la Recherche Scientifique Médicale, Belgium.

The community of the self

Timothy G. Buchman

Washington University School of Medicine, Department of Surgery, Box 8109, 660 South Euclid Avenue, St Louis, Missouri 63110-1093, USA
(e-mail: buchman@msnotes.wustl.edu)

Good health, which reflects the harmonious integration of molecules, cells, tissues and organs, is dynamically stable: when displaced by disease, compensation and correction are common, even without medical care. Physiology and computational biology now suggest that healthy dynamic stability arises through the combination of specific feedback mechanisms and spontaneous properties of interconnected networks. Today's physicians are already testing to 'see if the network is right'; tomorrow's physicians may well use therapies to 'make the network right'.

Claude Bernard, the father of physiology, observed that diverse physiological mechanisms had a common purpose: those mechanisms maintain the interior of a biological entity stable in the face of stress. The brilliance of Bernard's insight cannot be denied. Whether a perturbation affects multiple organ systems globally or is targeted to a specific organelle, restorative mechanisms have been observed and proven to be functionally important to physiological stability. More recently, mathematicians and physicists studying network models of biological systems have suggested that stability need not be engineered, but rather can emerge as a property of the network and its interconnections. At all levels — from genes to the web of organ systems that make up an individual — it is the balance of autonomy and connectedness that sustains health. These two founts of stability have complementary roles in guarding the communities of cells that, in aggregate, is the organism itself.

Historical background

Early in the twentieth century, the Harvard physiologist Walter B. Cannon embraced and extended Bernard's concept to suggest that tight regulation of physiological parameters — what he termed "homeostasis" — was the result of restorative mechanisms that both sensed and corrected deviations from a normal state (Fig. 1). A medical doctor by training, Cannon went on to suggest that failed homeostatic mechanisms precipitated illness, and that the purpose of the clinician was to substitute for those mechanisms until control could be restored¹. Cannon's imperative continues to drive the behaviour of physicians even today: the management of many common disease processes, such as diabetes mellitus, hypertension and hypercholesterolaemia, involves titration of drugs to normalize a measured value.

This substitution strategy seems most helpful when the disease process has a recognized mechanism and the drug(s) target the failure. More complex disease processes, particularly those that affect multiple mechanisms, have unpleasantly surprised well-intended doctors who have attempted to 'fix' parameters on their patients' behalf. For example, aggressive resuscitation to normalize blood pressure in penetrating trauma, infusion of calcium to correct the hypocalcaemia of sepsis, and (most recently) hormone therapies to replace menopausal deficiencies all may be harmful. Yet a talismanic belief in 'normal values' continues to permeate medical teaching and their restoration drives western medical practice.

A contemporary of Cannon's at Harvard University, Lawrence J. Henderson, embraced a more integrated perspective of biological regulation. His investigative focus was blood chemistry. Reasoning from physical and chemical principles of the time, it was Henderson who divined interrelationships from vast amounts of tabular data without benefit of computers. Henderson could predict numerically the behaviour of blood buffers provided that two values among the seven unknowns (total oxygen, total carbon dioxide, pO_2 , pCO_2 , pH, cell volume and anion concentration ratio across cell membranes) were presented. He developed nomograms that conveniently described those relationships and that were used subsequently by physicians for decades. Although students of biochemistry know him best for the eponymous Henderson–Hasselbach equation, his greater contribution may have been to stimulate physicians and scientists to evaluate physiological processes in the larger context of other biological systems. For Henderson, the organization of those systems and the mechanisms were not exclusive, but rather interdependent² (Fig. 1).

Systems theory and biology

Cannon and his disciples successfully pursued reductionist explanations of physiological stability in the face of perturbation. Pursuit, identification and description of mechanisms that restore and maintain measurable parameters ranging from vital signs (for example, temperature and blood pressure) to ion gradients remain a major research focus.

The notion that membership in a network could confer stability emerged from Ludwig von Bertalanffy's description of general systems theory in the 1930s and Norbert Wiener's description of cybernetics in the 1940s. General systems theory focused in part on the notion of flow, postulating the existence and significance of flow equilibria. In contrast to Cannon's concept that mechanisms should yield homeostasis, general systems theory invited biologists to consider an alternative model of homeodynamics in which nonlinear, non-equilibrium processes could provide stability if not constancy. What Wiener did was blend systems theory and control theory (embedding in the title of his masterwork the notion that his ideas were equally applicable to animals and to machines) and demonstrating that communication and control were inseparable.

The fusion of systems theory with biology into systems biology required two types of developments: the theory had to be adapted to biological systems, and biological systems had to produce data suitable for evaluation and testing³. Both have depended heavily on computation. In the theoretical domain, Haken recognized that elements in a system

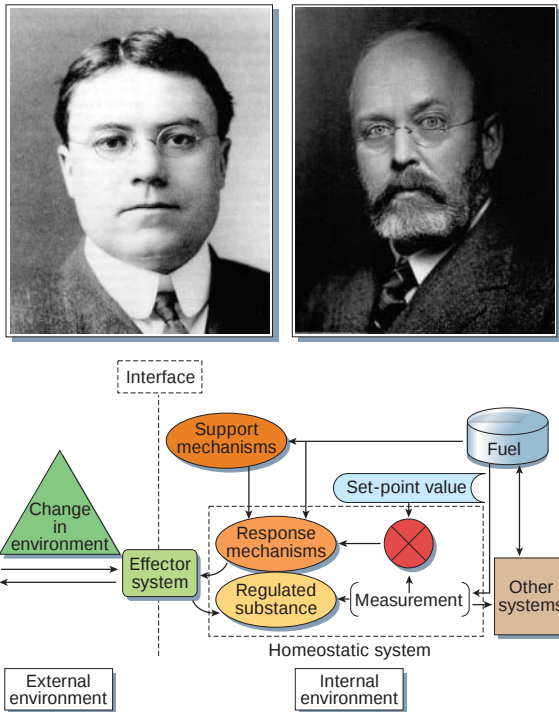


Figure 1 The concept of homeostasis. The top images are of Walter B. Cannon (left), who suggested that the presence of stable measurable parameters implied the presence of homeostatic systems, and his contemporary Lawrence J. Henderson (right), who provided a counterexample through his study of the buffering capacity of blood — the interactions among components were sufficient to confer stability without a separate regulating mechanism. The image below is a schematic of homeostasis. Changes in the environment are transduced to cause a change in the level of a regulated substance. This change is detected through measurement and comparison with a coded set-point value. Disparities between the measured value and the set-point value regulate a response mechanism that directly or indirectly influences effector systems at the exterior–interior interface. Homeostatic systems often require fuel, other support mechanisms and interact with other systems.

could cooperate spontaneously and developed rules about the factors that could facilitate and disrupt the appearance of coherent movement. The development of mathematical criteria that discriminate the spontaneous appearance of such cooperative behaviour led Haken to introduce the term ‘synergetics’. More important than the term is the idea that in non-equilibrium systems (such as networks whose integrity depends on the availability of metabolic fuels), stable states may be far apart from one another, and moreover that the system can jump from one stable state to another.

Kauffman’s computed simulation studies of toy systems configured as genetic networks showed that specifying the number and strength of connections among the elements of the network was sufficient to predict the canonical behaviours of the system (for example, the number of stable states). In aggregate, these experiments predicted that stability would arise in nearly every system in which elements that could assume a range of values were interconnected. In Kauffman’s words, the stability need not be engineered — rather, interconnected networks provide “order for free”⁴. Collectively, the work of Haken and of Kauffman formed a foundation from which research on network dynamics and modern gene regulatory networks evolved. Studies of real genetic networks (through transcriptome analysis of cells in diverse states) tend to support the prediction: large subsets of genes are often strongly

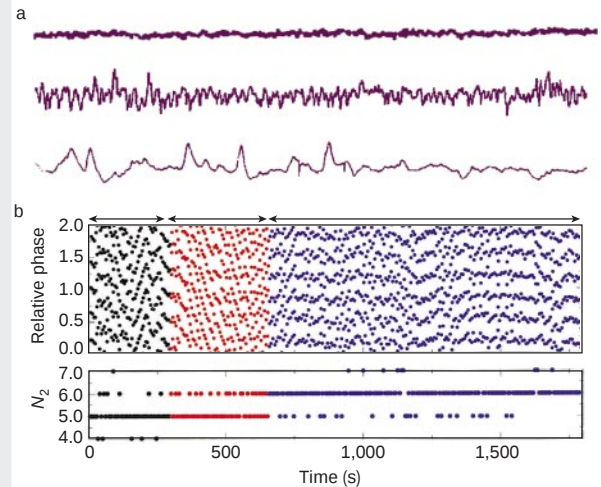


Figure 2 Interbeat variability in health, ageing and disease. **a**, Interbeat intervals obtained from electrocardiograms of three subjects (composite from refs 23, 35). Upper tracing, patient with severe congestive heart failure; middle tracing, healthy young subject; lower tracing, healthy aged subject. Compared to the healthy young subject, the aged subject and the patient with system failure have both lost variability, although the healthy aged subject retains more. **b**, A possible contributor to the healthy interbeat variability. This cardiorespiratory synchrogram illustrates the weak interaction between the cardiac and respiratory systems in a healthy young athlete at rest (image reproduced from ref. 36). The upper panel shows the phase relationship of each heartbeat within two respiratory cycles. The lower panel replots the data, showing that the coupling shifts from 5 heartbeats within 2 respiratory cycles ($N_2 = 5$) to 6 heartbeats within 2 respiratory cycles ($N_2 = 6$). Colours in the upper and lower panels show the transition (red) from 5:2 frequency locking (black) to 3:1 phase locking (blue). Studies by Goldberger and colleagues show that the seemingly ‘regular’ heartbeat of the young healthy heart is actually rather variable. Although some of that variability is accounted for by ventilation-dependent changes in the volume of blood returned to the heart, other sources of variation include rebalancing of autonomic tone. Both ageing and disease can increase the relative influence of sympathetic over parasympathetic components, leading to loss of variability. The synchrogram data suggest that uncoupling and recoupling are part of normal physiology. The acutely decoupled heart (such as the transplanted heart immediately following implantation) also has markedly diminished variability.

induced or repressed in an all-or-none response as cells move from one state to another.

Although science presently focuses on the simplest tractable system — individual cells — organization and regulation need to be defined at all levels of resolution if the source of stability is to be elucidated. Chauvet, studying mathematical models of formal biological systems, showed that nesting systems (biologically, this corresponds to the associations of cells into tissues, tissues into organs, and organs into the intact organism) could be spontaneous and would provide additional stability to the community of the self⁵. Kitano, who is one of the strongest proponents of modern systems biology, suggests that every system (irrespective of the level of resolution) be analysed with respect to the system’s structure — its dynamics, its method of control and its method of design^{6,7}.

Coordinate responses to external stress

The responses to stress are critical to the maintenance of the community of the self. Such responses are triggered by violation of structural integrity (trauma), failure of other regulatory systems (for example, skin barriers) that permit microbial invasion or replication, and even endogenous threats such as cancer. There are at least two systems requirements. First, the multiple responses must coordinate, and second, they must be contained in space and time.

Box 1

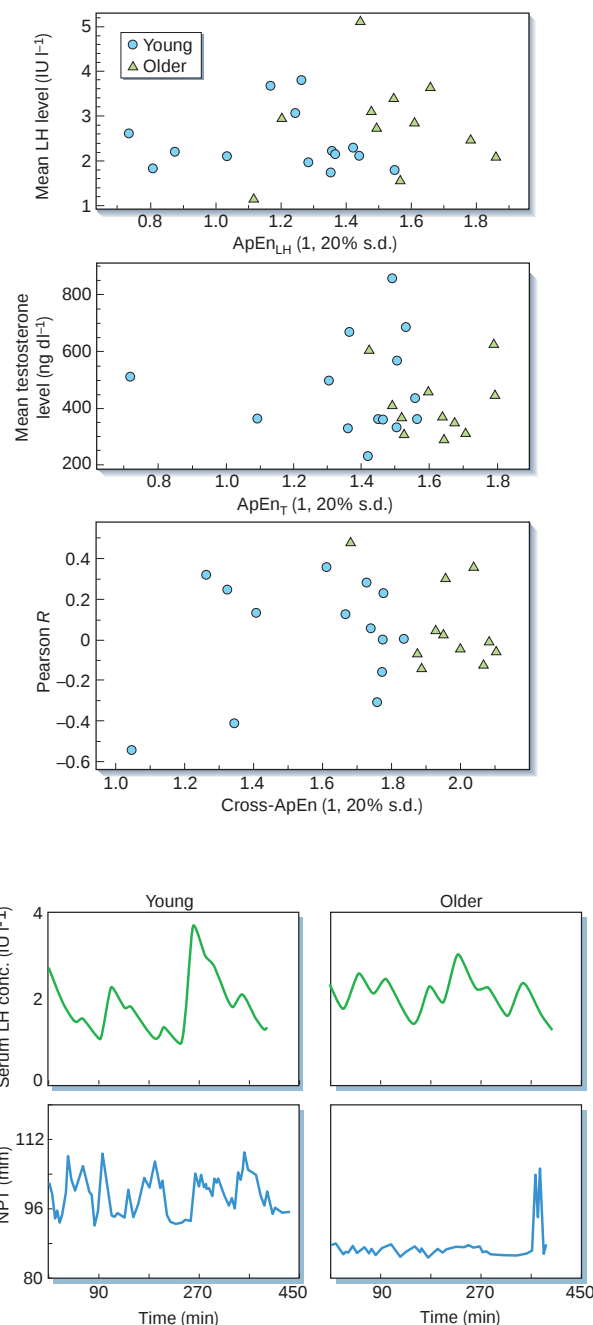
Tools to estimate coupling among systems

A task that challenges mathematical biologists who study the community of the self is the generation of unbiased estimates of coupling among different systems. One such estimate comes from a family of statistics called cross-ApEn. The parent family of statistics, approximate entropy or ApEn, measures the log likelihood that given each short pattern of a single data type, the next datum falls within an arbitrarily narrow range. Cross-ApEn measures the likelihood that given each short pattern of data in one data type, the next datum of another (putatively coupled) type falls within an arbitrarily narrow range. Two input parameters, m and r , are specified to compute ApEn and cross-ApEn. m represents the window (or vector) length of consecutive measurements; r represents the tolerance for testing sub-pattern regularity. In order to maintain scale invariance, r is conventionally defined as a percentage of the between-sample variation (for example, 20% s.d.).

Pincus, who developed the ApEn and cross-ApEn statistics, has applied them to several clinical situations. For example, a study was performed to determine possible secretory irregularities with ageing in the luteinizing hormone/testosterone (LH–T) secretory axis (see upper three figure panels opposite, redrawn with permission from ref. 25). Serum concentrations were derived for the two hormones in 14 young (aged 21–34 years) and 11 older (aged 62–74 years) healthy men. For each subject, blood samples were obtained at frequent (2.5-min) intervals during a sleep period. When the age contrast in ApEn values for the luteinizing hormone and testosterone time series were considered singly (top two panels), mean (and standard deviation) concentrations of the two hormones were indistinguishable in the two age groups. Visual inspection of the scatterplots suggest that the secretion of luteinizing hormone and testosterone were more regular (lower ApEn) in the young subjects; however the separation between young and old subjects is incomplete. But when cross-ApEn was applied to the paired LH–T time series (lower panel), older subjects exhibited greater cross-ApEn values (1.961 ± 0.121) compared to younger subjects (1.574 ± 0.249 ; $P < 10^{-4}$), with nearly 100% sensitivity and specificity, indicating greater LH–T asynchrony in the older group. Moreover, no significant differences in LH–T linear correlation (Pearson R ; $P > 0.6$) were found between the younger and older cohorts. Simple linear correlation does not detect the age-dependent differences in entwinement between secretion of luteinizing hormone and testosterone. Mechanistically, the results implicate LH–T network uncoupling as marking male reproductive ageing.

In a related study (see lower figure panels opposite, redrawn schematically from ref. 37, with permission), differences between younger and older males were studied based on several sex hormones and nocturnal penile tumescence (NPT) time series. The most vivid differences between the younger and older cohorts were that the paired LH–NPT dynamics were much more asynchronous in the older group. The timing between regulated hormonal (LH) input and target sexual response (NPT) output was significantly disrupted in the ageing male. This reinforces the point that successful therapeutic strategies probably need to be more integrative or network-oriented, rather than strictly local in structure.

Tools such as cross-ApEn facilitate pairwise comparison between the kind of time-series data that can be obtained, for example, from whole organs. Analysis of massively parallel data sets, such as data obtained from genomic studies, are hindered by high dimensionality:



the inter-relationships among $\sim 10^4$ data types are difficult to define from $\sim 10^1$ experiments (a value typical of current genomic experimental design). An important research goal would therefore seem to be the development of new tools that could extract possible inter-relationships among the data types and order the probability of those relationships objectively.

The main stress responses occur within cells and among cells. The principal intracellular response, originally called the heat-shock response, revises transcription and reassigns translation to produce several dozen gene products ('stress proteins') that collectively refocus metabolism on cleaning up denatured structures and fortifying

the cell (at least temporarily) against subsequent insults. All cells produce the major heat-shock protein, Hsp72, an inducible chaperonin that sequesters and refolds damaged proteins (ref. 8; and see review in this issue by Koonin *et al.*, pages 218–223). The quantity and persistence of Hsp72 (and the other stress proteins) is roughly

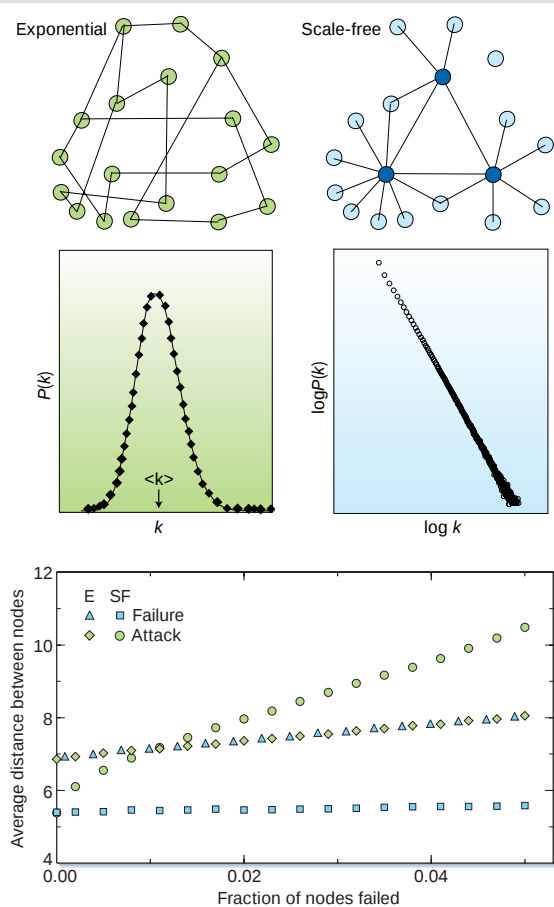
Box 2

Network structure affects tolerance to node failure

Networks seem to promote physiological stability. Barabasi's group has compared two canonical types of network — exponential networks (also known as Erdős–Rényi networks) and scale-free networks (see upper four figure panels opposite, reproduced from ref. 38). The connectivities of these networks are different and are characterized by the probability $P(k)$ that a given node has k links. For an exponential network, $P(k)$ peaks strongly at $k = \langle k \rangle$ and decays exponentially for large k . In the scale-free network, most nodes have only a few links, but a few nodes, called hubs (dark blue), have a very large number of links. In this case $P(k)$ has no well-defined peak, and for large k it decays as a power-law, appearing as a straight line on a log–log plot.

Natural metabolic networks seem to be mostly of the scale-free type. This architecture carries specific implications concerning node dysfunction and consequent network failure. Provided that nodes fail randomly, interconnectedness is far better preserved in scale-free networks than in exponential networks. However, if a disease process attacks key nodes, the scale-free network is more susceptible to failure (see lower figure panel opposite, from ref. 39). Distances between nodes are, on average, smaller in scale-free (SF) networks than in exponential (E) networks that connect equal numbers of nodes (see distance at zero failure). As the fraction of failed nodes increases, the residual connectedness depends on the underlying architecture and whether the failure is random or occurs by attack on hubs.

Once the network has failed, restoration of function to key affected nodes is necessary, but the appropriate connections must be restored to ensure resurrection of function. In general, aged patients recover from serious illness much more slowly than do younger, similarly ill patients. It is possible that aged patients cannot search through the space of possible connections as efficiently as their younger counterparts. If so, then mathematical biologists may be able to help clinicians by exploring treatment strategies that guide the search as opposed to fixing the value of particular nodes.



proportional to the intensity and duration of the stress. Resolution of the stress is sufficient to attenuate the heat-shock response — no anti-heat-shock programme has been identified — and return the affected cell to its basal programmes of gene expression and metabolic function.

In contrast, inflammation is a cooperative response involving multiple cell types, orchestrated both locally and remotely, and affecting the host at multiple levels of resolution (from organism to gene expression). Extracellular responses depend on biological amplification, and so differ from the proportionate intracellular response. Following an ordinary laceration — such as a nick or cut caused by a razor — the dynamics of the response are familiar if not entirely predictable. Priority is given to controlling the associated haemorrhage as platelets are recruited to the injury site to form a plug, and the coagulation cascade is activated. Chemical signals are emitted from the wound, which alter both blood flow and vessel characteristics. The latter alterations also cause circulating cells such as polymorphonuclear leukocytes (PMNs) to stick to the endothelium lining the blood vessel wall and to traverse it. Once sequestered into the response site, PMNs inflict oxidative damage upon foreign entities and mononuclear phagocytic cells are recruited to clean up the debris. If the clean up is successful, the response fades away. Alternatively, if microbes become established and begin to proliferate, the inflammatory response widens to block the invaders through abscess formation.

Unlike the intracellular stress response, this multicellular inflammatory response involves multiple signal amplifications. Specific containment strategies are necessary to keep the processes in

check. Local clotting does not become widespread because circulating anticoagulants (for example, protein C) are activated. Thrombus formation also invites activation of fibrinolysis. PMNs recruited to the site will die by apoptosis. Even if apoptosis is delayed, an internal switch in lipid biosynthesis (from leukotrienes and prostaglandins to lipoxins) caused by the interaction of PMNs with cells resident in the inflamed tissue reprogrammes those PMNs to promote resolution of inflammation⁹. The mononuclear cells and T-helper lymphocytes that initially secrete pro-inflammatory cytokine molecules shift synthetic programmes to favour anti-inflammatory cytokines. In isolation, these containment strategies also seem to fit reductionist models. The strategies agree perfectly with Cannon's construct of homeostasis, and the medical response to failure of homeostasis directly supports them: if the patient is still bleeding, administer procoagulants, and if the patient is clotting excessively, prescribe blood thinners.

However, the processes are not isolated. In ordinary parallel operation, the containment aspects of the intracellular and multicellular responses become decidedly nonlinear. This manifests clinically as context dependence of the response. Matzinger has suggested and demonstrated that the immune system does not merely distinguish between self and non-self; rather, it responds to new epitopes only when there is a concurrent signal indicating that the organism is endangered¹⁰. The consequence of inducing a heat-shock response in endothelial cells and in fibroblasts is similarly context dependent. When an inducer of the heat-shock response is applied to cells in their normal state, the cells become refractory to the adverse consequences of a subsequent inflammatory stimulus such as bacterial

lipopolysaccharide. In contrast, cells that have been recently stimulated with bacterial lipopolysaccharide and are then exposed to an inducer of the heat-shock response execute apoptosis¹¹. In the language of algebra, the operators of the stress response, inflammation and heat shock do not commute (see review in this issue by Searls, pages 211–217). Thrombosis prior to an inflammatory response is typically self-limited. Thrombosis in the context of an inflammatory response can lead to regional or even disseminated intravascular coagulation, probably through failure of the containment mechanisms. Conventional negative feedback models of homeostasis do not readily explain such context dependencies.

Some have suggested that the context dependencies are merely artefacts of experimentation and of modern therapy, that is, maladaptive responses to the invention of the hollow needle, intravenous fluids, antibiotics and other paraphernalia of intensive care. Given the recent discovery of hardwired safety mechanisms, this seems unlikely. For example, Tracey and colleagues have recently identified a vagally mediated circuit through which the brain can directly shut down the production of inflammatory mediators by the largest population of fixed mononuclear cells in the human body^{12,13}. Such safety mechanisms suggest that the context-dependent nonlinear interactions are necessary to the maintenance of the community of the self. The nonlinearities may eventually translate into therapeutic strategies. Weiss and colleagues have shown that viral transfer of the gene coding for Hsp70 into the lungs of rats markedly attenuates the pulmonary component of widespread inflammation initiated by bacterial peritonitis^{14,15}.

Multiple organ dysfunction syndrome

Safety mechanisms do sometimes fail. Until recently, the disruption of the community of the self following such failures precipitated death. Modern critical care, capable of supporting or even temporarily replacing the physiological function of vital organs, spawned a new disease, the multiple organ dysfunction syndrome (MODS). MODS is characterized by unbridled inflammation, remote in space and in time from the inciting event. Typically, an infection that seems to have been promptly identified and appropriately treated nevertheless causes a body-wide inflammatory response, leading to serial failure of the respiratory system, digestive system and renal system to perform their vital functions. The mortality of three-organ dysfunction is 60–80%, and MODS remains a leading cause of death in intensive care units. It now seems possible that MODS is the manifestation of widespread network failure^{16–19}.

Workers observed that during the descent into MODS, physiological time signals (such as the beat-to-beat interval of the electrocardiogram) would lose the fine variability observed in healthy patients. Reasoning from Pincus's observation that greater regularity could indicate increased system isolation, Godin suggested that unbridled inflammation could cause uncoupling of organs from one another, thus precipitating MODS^{18,19}. Godin went on to show that injection of bacterial endotoxin into human volunteers seemed to cause mild uncoupling that manifest as loss of variability in the electrocardiogram signal²⁰. Goldstein and colleagues have observed similar uncoupling of autonomic regulation in patients descending into clinical septic shock, and recoupling during the recovery period²¹. Neither is the uncoupling phenomenon peculiar to sepsis — it is also observed in the context of severe brain injury²².

Although such observations do not prove unequivocally that network disruption is the cause of MODS, the clinical imperative would seem to include restoration and protection of network integrity. Rebuilding a network is likely to be an orderly process; a common sequence of organ failure — lungs, then gut/liver, and finally kidneys — is precisely the opposite order in which those organs mature during fetal life. Because of this, it could well be counterproductive to adopt the usual therapeutic strategy of adjusting and clamping measured outcomes (such as blood pressure and

pH) that serve as proxies of organ function simultaneously into their normal ranges.

Connectedness and mechanisms

It now seems that Cannon and Henderson were both correct. The community of the self appears to depend on classical homeostatic mechanisms as well as network integrity. An observed disruption of the stable state now raises a critical, if generic, issue: is it a specific mechanism or is it the connectedness among components that underlies the disruption? Mathematical biologists may help provide answers suggested by studies of physiological time signals in normal ageing (Fig. 2). Aged patients have stable, yet brittle, physiology. In addition to diminished reserves in nearly all systems, aged patients recover only slowly — or not at all — once physiological reserves are exceeded and medical intervention is necessary. It is not that aged patients have maladaptive responses to stress — rather their adaptive responses are inadequate. Goldberger and colleagues, reviewing fractal dynamics in physiology, note that many physiological time signals — from the heartbeat to gait — exhibit long-range correlation and rich multiscale dynamics. With age, such signals break down in two general ways, either exhibiting excessive order or uncorrelated randomness. These patterns of breakdown suggest that, for any physiological system, there is a range of connectedness that is optimal²³.

There exist several families of statistics that describe the regularity of time signals. One of them, approximate entropy (ApEn) serves as a useful example (Box 1). ApEn describes the likelihood that patterns recur in time signals²⁴ — decreases in ApEn correspond to decreases in pattern variability. The time signal of the heart (that is, the beat-to-beat intervals of the electrocardiogram) loses variability in normal ageing as it does in MODS. This loss of ApEn can be interpreted as isolation of the time signal source from the network. A derivative of ApEn called cross-ApEn describes the synchrony (coupling) that exists between pairs of time signals²⁵. When coupling is tight, events in one signal cause predictable patterns in the other signal and cross-ApEn remains low. Tightly coupled time signals are common in classical homeostatic mechanisms owing to the link between signals and responses. An example of such linked signals is luteinizing hormone and testosterone in males. With ageing, the relationship between luteinizing hormone and testosterone becomes asynchronous; this manifests as an increased cross-ApEn value, and suggests that the tight link has begun to erode. Statistics such as ApEn and cross-ApEn may prove more generally useful to interrogate physiological time signals for evidence of mechanistic integrity and network integrity.

Although network integrity is important to health, it is not a strict proxy. Schäfer and colleagues, studying the weak interaction between the cardiac and respiratory systems in healthy athletes at rest, noted that the two systems uncouple and recouple every few minutes²⁶ (Fig. 2). The connection strengths between systems, and probably within systems, seem to be plastic. Kaneko's laboratory recently reported that even such plastic networks composed of chaotic units will self-organize into hierarchical structures²⁷. Provided that the plasticity is limited, such uncoupling and recoupling may be part of healthy network physiology at many scales of resolution, from gene networks to organ systems. In this perspective, MODS may be less a problem of extended uncoupling and more a problem of failure to recouple (Box 2).

Such coupling relationships span multiple levels of resolution, an observation that demands bridging of molecular mechanisms and genetic data with physiological systems and function. Noble's wry vision of "genes as 'prisoners' that are trapped inside the successful physiological systems that express them" speaks to the ambiguity of causation in complex networks²⁸. Although bench studies of network relationships provide important estimates of kinetics, their scope is necessarily narrow and confined typically to a single level. To address this limitation, projects such as the Human Physiome Project (<http://www.physiome.org.nz/anatml/pages/index.html>)

are undertaking development of tools and computer languages that capture representations of known pathways and convert them to common modelling code, thus facilitating merger and enabling a more complete visualization.

This approach has yielded substantial progress in modelling and understanding the human heart. Here, parallel research tracks in genetics, cell biology and organ physiology have been linked through Physiome Project tools to provide new insight into normal pathways of electrical conduction and the pathological events that are clinically manifested as life-threatening arrhythmias²⁹. Finer-grained but more richly detailed simulations of intracellular events are being performed using the Virtual Cell modelling environment (http://www.nrcam.uchc.edu/vcell_development/vcell_dev.html), describing network-dependent phenomena such as calcium oscillations³⁰ and nuclear envelope breakdown.

As made apparent by Pincus, tighter coupling of independent signals with random components increases the stochastic behaviour of both. Tightening the coupling and increasing the stochastic components may have a physiological advantage. Collins and colleagues have explored the consequences of applied stochastic resonance, which arises from superposition of different types of noise (white, power-law and even coloured noise with long-range correlation) upon a low-amplitude signal. The noise serves to transiently amplify the signal above an otherwise unattained threshold. An exogenous noisy background thereby improves such diverse functions as tactile perception in diabetic neuropathy and perhaps even the effectiveness of life-supporting mechanical ventilation^{31–34}. Perhaps more important, such noise appears endogenously in several physiological systems (see review in this issue by Rao and colleagues, pages 231–237).

Systems physiology: back to the future?

During the heyday of systems physiology — the middle third of the twentieth century — the inventory of many biological laboratories included larger animals, pressure transducers, flow probes and strip chart recorders. Relationships among data were distilled exclusively through inspection, and the analytic tool of the day (regression) concealed a bias for reductionist models that are now recognized to inadequately describe interconnected networks. The dawn of genomics (and, more generally, the promises of molecular medicine) tumbled systems physiology as a leading investigative approach. Yet clinicians continue to rely most heavily on systems physiology — such as bedside haemodynamics and respiratory dynamics — to promote the integrity of self, suggesting that a deeper understanding may yield new therapies.

Interest is now being rekindled in studies of systems physiology, especially those conducted in concert with genomic, transcriptomic, proteomic and metabolomic investigations. Digitized streams of physiological parameters create new analytic challenges that can be met only through partnerships among theorists, experimentalists and analysts. It is vital to create models that embed homeostatic mechanisms into larger networks that themselves confer robustness to perturbation and thereby protect the community of the self. But more important, and much harder, will be determining whether a particular model or class of models properly captures the protective behaviours reflected across multiple resolutions, from genes to humans. □

doi:10.1038/nature01260

- Chambers, N. K. & Buchman, T. G. Shock at the millennium. I. Walter B. Cannon and Alfred Blalock. *Shock* **13**, 497–504 (2000).
- Chambers, N. K. & Buchman, T. G. Shock at the millennium II. Walter B. Cannon and Lawrence J. Henderson. *Shock* **16**, 278–284 (2001).

- Wolkenhauer, O. Systems biology: the reincarnation of systems theory applied in biology? *Brief Bioinform.* **2**, 258–270 (2001).
- Kauffman, S. A. *The Origins of Order: Self-Organization and Selection in Evolution* (Oxford Univ. Press, Oxford, 1993).
- Chauvet, G. A. Hierarchical functional organization of formal biological systems: a dynamical approach. I. The increase of complexity by self-association increases the domain of stability of a biological system. *Phil. Trans. R. Soc. Lond. B* **339**, 425–444 (1993).
- Kitano, H. in *Foundations of Systems Biology* (ed. Kitano, H.) 1–36 (MIT Press, Cambridge, MA, 2001).
- Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662–1664 (2002).
- Hartl, F. U. & Hayer-Hartl, M. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* **295**, 1852–1858 (2002).
- Levy, B. D., Clish, C. B., Schmidt, B., Gronert, K. & Serhan, C. Lipid mediator class switching during acute inflammation: signals in resolution. *Nature Immunol.* **2**, 612–619 (2001).
- Matzinger, P. The danger model: a renewed sense of self. *Science* **296**, 301–305 (2002).
- DeMeester, S. L., Buchman, T. G. & Cobb, J. P. The heat shock paradox: does NF- κ B determine cell fate? *FASEB J.* **15**, 270–274 (2001).
- Bernik, T. R. et al. Pharmacological stimulation of the anti-inflammatory pathway. *J. Exp. Med.* **195**, 781–788 (2002).
- Blalock, J. E. Harnessing a neural-immune circuit to control inflammation and shock. *J. Exp. Med.* **195**, F25–F28 (2002).
- Weiss, Y. G., Maloyan, A., Tazelaar, J., Raj, N. & Deutschman, C. S. Adenoviral transfer of HSP-70 into pulmonary epithelium ameliorates experimental acute respiratory distress syndrome. *J. Clin. Invest.* **110**, 801–806 (2002).
- Slutsky, A. S. Hot new therapy for sepsis and the acute respiratory distress syndrome. *J. Clin. Invest.* **110**, 737–739 (2002).
- Marshall, J. C. Inflammation, coagulopathy, and the pathogenesis of multiple organ dysfunction syndrome. *Crit. Care Med.* **29**(Suppl.), S99–S106 (2001).
- Seely, A. J. & Christou, N. V. Multiple organ dysfunction syndrome: exploring the paradigm of complex nonlinear systems. *Crit. Care Med.* **28**, 2193–2200 (2000).
- Pincus, S. M. Greater signal regularity may indicate increased system isolation. *Math. Biosci.* **122**, 161–181 (1994).
- Godin, P. J. & Buchman, T. G. Uncoupling of biological oscillators: a complementary hypothesis concerning the pathogenesis of multiple organ dysfunction syndrome. *Crit. Care Med.* **24**, 1107–1116 (1996).
- Godin, P. J. et al. Experimental human endotoxemia increases cardiac regularity: results from a prospective, randomized, crossover trial. *Crit. Care Med.* **24**, 1117–1124 (1996).
- Ellenby, M. S. et al. Uncoupling and recoupling of autonomic regulation of the heart beat in pediatric septic shock. *Shock* **16**, 274–277 (2001).
- Goldstein, B., Towell, D., Lai, S., Sonnenthal, K. & Kimberly, B. Uncoupling of the autonomic and cardiovascular systems in acute brain injury. *Am. J. Physiol.* **275**, R1287–R1292 (1998).
- Goldberger, A. L. et al. Fractal dynamics in physiology: alterations with disease and aging. *Proc. Natl Acad. Sci. USA* **99**, 2466–2472 (2002).
- Pincus, S. M. Approximate entropy as a measure of system complexity. *Proc. Natl Acad. Sci. USA* **88**, 2297–2301 (1991).
- Pincus, S. M. et al. Older males secrete luteinizing hormone (LH) and testosterone more irregularly, and jointly more asynchronously, than younger males. *Proc. Natl Acad. Sci. USA* **93**, 14100–14105 (1996).
- Schäfer, C., Rosenblum, M. G., Abel, H.-H. & Kurths, J. Synchronization in human cardiorespiratory system. *Phys. Rev. E* **60**, 657–670 (1999).
- Ito, J. & Kaneko, K. Spontaneous structure formation in a network of chaotic units with variable connection strengths. *Phys. Rev. Lett.* **88**, 028701–028701–4 (2002).
- Noble, D. The rise of computational biology. *Nature Rev. Mol. Cell Biol.* **3**, 460–462 (2002).
- Noble, D. Modeling the heart—from genes to cells to the whole organ. *Science* **295**, 1678–1682 (2002).
- Fink, C. C. et al. An image-based model of calcium waves in differentiated neuroblastoma cells. *Biophys. J.* **79**, 163–183 (2000).
- Collins, J. J., Chow, C. C., Imhoff, T. T. Aperiodic stochastic resonance in excitable systems. *Phys. Rev. E* **52**, R3321–R3324 (1995).
- Suki, B. et al. Life-support system benefits from noise. *Nature* **393**, 127–128 (1998).
- Neiman, A., Schimansky-Geier, L., Moss, F., Shulgin, B. & Collins, J. J. Synchronization of noisy systems by stochastic signals. *Phys. Rev. E* **60**, 284–292 (1999).
- Gong, Y., Matthews, N. & Qian, N. Model for stochastic-resonance-type behavior in sensory perception. *Phys. Rev. E* **65**, 031904–031904–5 (2002).
- Lipsitz, L. A. & Goldberger, A. L. Loss of ‘complexity’ and aging. Potential applications of fractals and chaos theory to senescence. *J. Am. Med. Assoc.* **267**, 1806–1809 (1992).
- Schäfer, C., Rosenblum, M. G., Kurths, J. & Abel, H. H. Heartbeat synchronized with ventilation. *Nature* **392**, 239–240 (1998).
- Veldhuis, J. D., Iranmanesh, A., Mulligan, T. & Pincus, S. M. Disruption of the young-adult synchrony between luteinizing hormone release and oscillations in follicle-stimulating hormone, prolactin, and nocturnal penile tumescence (NPT) in healthy older men. *J. Clin. Endocrinol. Metab.* **84**, 3498–3505 (1999).
- Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A. L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
- Albert, R., Jeong, H. & Barabási, A. L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).

Acknowledgements

Work in the author's laboratory is supported by the National Institute of General Medical Sciences and the National Institute of Nursing Research. T.G.B. thanks the following colleagues for their comments during preparation and review of this article: N. Chambers, P. Cobb, J. Collins, W. Fontana, B. Goldstein, K. Kaneko, H. Kitano, D. Noble, S. Pincus and K. Tracey.

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Senior European Sales Manager:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Netherlands/ Italy/ Iberia:

Evelina Rubio Hakansson (4973)

Scandinavia: Silje Opstrup (4994)

France/ Belgium:

Amelie Pequignot (4974)

Production Manager: Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Hind Berrada (4935)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria/ Switzerland:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Harajukamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

e-mail: k.johnson@naturejpn.com

Asia-Pacific Advertising Manager:

Keyn Johnson

naturejobs

Part-time growth

Job opportunities within academia have increased rapidly over the past 20 years — at least, for people willing to work part-time. According to the American Council on Education, which recently analysed data from the US Department of Education, the number of part-time faculty members in the United States rose by 79% between 1981 and 1999, to more than 400,000 out of a total of one million instructors. The biggest rise occurred between 1987 and 1992, when both the economy and the college-age population were growing. During that period, 82% of the 120,000 new faculty members who were hired took up part-time positions.

The analysis of that analysis? For science and medicine, the situation could have been worse. The proportion of part-time positions in health-sciences faculties grew from 42.8% in 1992 to 46.6% in 1998, with the balance of tenure and tenure-track positions shifting from 36% in 1992 to 32.7% in 1998. In natural sciences, the percentage of part-time positions fell during the same period, from 33% in 1992 to 32.2% in 1998. But the proportion of full-time, non-tenure-track positions almost doubled, from 3.7% to 7.3%.

The story is similarly mixed when looking at different kinds of institution. Private research universities reduced their dependence on part-timers, from 46% of faculty members in 1992 to 40% in 1998. The opposite was true for public research institutions, where the percentage of part-time positions rose from 26% to 30%.

So how can you improve your odds of securing full-time work? One way is to mirror the part-timers who have several jobs and publish almost as much as their full-time colleagues in hope of getting on the tenure track. Another is to increase the odds, but reduce long-term security, by considering full-time, non-tenure-track positions, which may be a growth area.

Paul Smaglik

Naturejobs editor



Contents

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS